



**HAL**  
open science

## Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition

Najwa Taib, Daniela Megrian, Jerzy Witwinowski, Panagiotis Adam, Daniel Poppleton, Guillaume Borrel, Christophe Beloin, Simonetta Gribaldo

► **To cite this version:**

Najwa Taib, Daniela Megrian, Jerzy Witwinowski, Panagiotis Adam, Daniel Poppleton, et al.. Genome-wide analysis of the Firmicutes illuminates the diderm/monoderm transition. *Nature Ecology & Evolution*, 2020, 4 (12), pp.1661-1672. 10.1038/s41559-020-01299-7 . pasteur-03207762

**HAL Id: pasteur-03207762**

**<https://pasteur.hal.science/pasteur-03207762v1>**

Submitted on 4 May 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

1 **Genome-wide analysis of the Firmicutes illuminates**  
2 **the diderm/monoderm transition**

3  
4 Najwa Taib<sup>1,2#</sup>, Daniela Megrian<sup>1,3#</sup>, Jerzy Witwinowski<sup>1</sup>, Panagiotis Adam<sup>1^</sup>, Daniel Poppleton<sup>1&</sup>,  
5 Guillaume Borrel<sup>1</sup>, Christophe Beloin<sup>4</sup>, and Simonetta Gribaldo<sup>1</sup>

6 <sup>1</sup> Unit Evolutionary Biology of the Microbial Cell, Department of Microbiology, Institut Pasteur, UMR  
7 2001 CNRS, Paris, France

8 <sup>2</sup> Hub Bioinformatics and Biostatistics, Department of Computational Biology, Institut Pasteur, USR  
9 3756 CNRS, Paris, France

10 <sup>3</sup> Sorbonne University, Collège doctoral, F-75005 Paris, France

11 <sup>4</sup> Unit Genetics of Biofilms, Department of Microbiology, Institut Pasteur, Paris, France

12 <sup>^</sup> Current affiliation: Environmental Microbiology and Biotechnology, Faculty of Chemistry,  
13 University Duisburg-Essen, Germany

14 <sup>&</sup> Current affiliation: Department of Comparative Biomedical Sciences, Royal Veterinary College,  
15 University of London, UK

16  
17 # these authors contributed equally to this work

18 \* Correspondence: [simonetta.gribaldo@pasteur.fr](mailto:simonetta.gribaldo@pasteur.fr)

19 Keywords: Limnochordia; Cell Envelope; Outer Membrane; LPS; Phylogenomics; Evolution; Diderm  
20 Firmicutes

21  
22 **Summary**

23 The transition between cell envelopes with one membrane (Gram-positive or monoderm) and those with  
24 two membranes (Gram-negative or diderm) is a fundamental open question in the evolution of Bacteria.  
25 The evidence of two independent diderm lineages, the Halanaerobiales and the Negativicutes, within  
26 the classically monoderm Firmicutes has blurred the monoderm/diderm divide and specifically  
27 anticipated that other members with an outer membrane (OM) might exist in this phylum. Here, by  
28 screening 1,639 genomes of uncultured Firmicutes for signatures of an OM, we highlight a third and  
29 deep branching diderm clade, the Limnochordia, strengthening the hypothesis of a diderm ancestor and  
30 multiple transitions leading to the monoderm phenotype. Phyletic patterns of over 176,000 protein  
31 families constituting the Firmicutes pan-proteome identify those that are specific to the three diderm  
32 lineages, and suggest new potential players in OM biogenesis. In contrast, we find practically no largely  
33 conserved core for monoderms, a fact possibly linked to different ways of adapting to OM loss.  
34 Phylogenetic analysis of a concatenation of main OM components totalling nearly 2000 amino acid

35 positions illustrates the common origin and vertical evolution of most diderm bacterial envelopes.  
36 Finally, mapping the presence/absence of OM markers onto the tree of Bacteria highlights the  
37 overwhelming presence of diderms and the non-monophyly of monoderms, pointing to an early origin  
38 of two-membraned cells and the derived nature of the Gram-positive envelope following independent  
39 OM losses.

40

## 41 **Introduction**

42 The cell envelope is one of the most essential and ancient features of life; yet, most aspects of its  
43 evolutionary history remain obscure. In particular, it is unclear how cell envelopes with two membranes  
44 (Gram-negative or diderms) and those with one membrane (Gram-positive or monoderms) came into  
45 being, and how such important transition occurred<sup>1-5</sup>.

46 The Firmicutes are the textbook example of Gram-positive bacteria, but surprisingly include  
47 two clades, the Negativicutes and the Halanaerobiales, that display an outer membrane (OM) with  
48 lipopolysaccharide (LPS)<sup>6-9</sup>. We recently showed that they form two phylogenetically distinct lineages,  
49 each close to different monoderm relatives within the Firmicutes. In addition, phylogenetic analysis of  
50 core biosynthetic LPS genes indicated that these were not acquired through horizontal gene transfer  
51 (HGT) from other diderm bacteria. Finally, identification and annotation of putative OM markers in  
52 the genomes of Halanaerobiales and Negativicutes suggested that these two lineages display unique cell  
53 envelopes with specific characteristics<sup>10</sup>, such as for example the mechanism of OM attachment, which  
54 is different from that of *Escherichia coli*<sup>5,11</sup>. These bioinformatics predictions were confirmed by  
55 characterizing the first OM proteome from the model diderm Firmicute *Veillonella parvula*<sup>12</sup>. From these  
56 results, we put forward the hypothesis that a diderm envelope with LPS was already present in the  
57 ancestor of all Firmicutes and was retained in the Negativicutes and Halanaerobiales while it was lost  
58 multiple times independently during the diversification of this phylum to give rise to the classical Gram-  
59 positive cell envelope architecture<sup>10</sup>. This hypothesis specifically anticipated that other diderm lineages  
60 may exist in the Firmicutes.

61 Recently, 1,639 new genomes from uncultured Firmicutes obtained from metagenomes were  
62 made available<sup>13</sup>, providing an exceptional wealth of new data to explore. Here, we analyzed these  
63 genomes to investigate the presence of additional diderm members and to obtain further information on  
64 the diderm/monoderm transition in this phylum. Our results strengthen the hypothesis of a diderm  
65 ancestor of the Firmicutes and possibly all Bacteria, and the derived nature of the monoderm cell  
66 envelope.

67

## 68 **Results**

### 69 *Screening uncultured Firmicutes genomes for OM markers highlights a third diderm lineage*

70 We retrieved 1,639 genomes annotated as Firmicutes from the Uncultured Bacteria and Archaea (UBA)  
71 dataset of Parks et al.,<sup>13</sup>. These genomes were isolated from different environments and their quality

72 goes from partial to near complete (Supplementary Table 1 and Supporting Data). 514 UBA genomes  
73 having less than 35 ribosomal proteins were considered as too partial and discarded from further  
74 analysis. To robustly place these new genomes in the reference Firmicutes phylogeny, we included the  
75 1,125 UBA genomes in a phylum-level tree together with 230 representatives of all Firmicutes families,  
76 and members of major bacterial phyla as outgroup (Methods).

77 The resulting maximum-likelihood (ML) tree is well resolved (Figure 1 and Supporting Data).  
78 The UBA genomes significantly enrich the genomic coverage of the Firmicutes, by spanning the entire  
79 diversity of this phylum, in particular the Clostridia. Most UBA genomes fall into known Firmicutes  
80 families consistently with their inferred taxonomy<sup>13</sup>, whereas other fall into clades that do not contain  
81 any genome representative of known families, and were given a loose taxonomic assignment<sup>13</sup> (Figure  
82 1). Only one UBA genome belongs to the Halanaerobiales, possibly a bias due to the difficulty of  
83 assembling these GC-rich genomes or a poorer sampling from the environments where they thrive. In  
84 contrast, these new genomes significantly enrich the coverage for Negativicutes (39 UBAs). The deep  
85 branching of Halanaerobiales and the placement of Negativicutes within Clostridia are both well  
86 supported and are consistent with previous analyses<sup>6,10,14</sup>. In particular, Antunes et al. specifically tested  
87 alternative branchings of the Halanaerobiales by approximately unbiased (AU) tests and showed that  
88 they were all strongly rejected by the data<sup>10</sup>. Interestingly, while the monoderm lineage *Natranaerobius*  
89 *thermophilus* branched with Halanaerobiales in previous analyses<sup>10,14</sup>, it now groups with *Dethiobacter*  
90 *alkaliphilus* and 17 other UBA genomes at the base of Bacilli (Figure 1). The increased genomic coverage  
91 likely enhanced the phylogenetic signal and helped to correctly place these lineages, which is more  
92 consistent with their monoderm phenotype.

93 To investigate the existence of additional diderm lineages among the UBA Firmicutes, we  
94 screened them for the presence of four conserved genes for LPS biosynthesis and six protein domains  
95 previously used as markers of Gram-negativity<sup>7</sup> (Methods). Consistently with our previous study<sup>10</sup>, we  
96 found homologues of these OM markers in all Negativicutes and Halanaerobiales UBA genomes, but  
97 surprisingly we also found them in 46 unclassified UBA genomes belonging to a particularly interesting  
98 clade (Supplementary Table 2). It represents the second deepest-branching group in the reference  
99 phylogeny after the Halanaerobiales (Figure 1), and includes a single isolated member, *Limnochorda pilosa*,  
100 identified from a brackish meromictic lake, and defining the recently proposed class Limnochordia<sup>15</sup>.  
101 The recently sequenced *L. pilosa* genome revealed the presence of classical OM markers consistent with  
102 microscopy evidence for a diderm cell envelope<sup>16</sup>. Most UBA genomes that we assign to Limnochordia  
103 were assembled from anaerobic mud and digester samples (Supplementary Table 1) and form a diverse  
104 group (Supplementary Figure 1). These results show that Limnochordia represent a third and deep-  
105 branching diderm lineage within the Firmicutes, strengthening the hypothesis of a diderm ancestor.

106

107 *Distribution of protein families and domains reveals the functional core of diderm Firmicutes*



108 The existence of three distinct diderm lineages embedded in the classical monoderm Firmicutes provides  
109 an exceptional opportunity to gain insights into the diderm/monoderm transition by analyzing the  
110 distribution of proteins across this phylum. To this aim, we assembled a reduced genome databank of  
111 316 representative Firmicutes (including the newly identified UBA diderm Firmicutes): 47  
112 Limnochordia, 101 Negativicutes and 17 Halanaerobiales, for a total of 165 diderm and 151 monoderm  
113 taxa (Methods). We used this databank to compare the proteomes of monoderm and diderm lineages  
114 across the phylum. We calculated the pan-proteome of Firmicutes as composed of 176,024 protein  
115 families, 15,964 being present in at least five genomes and which were thus kept for further analysis  
116 (Methods).

117 We then built a presence/absence matrix and calculated the distribution of these 15,964 protein  
118 families (Figure 2). Some families are specific to Halanaerobiales (327), to Negativicutes (1,820), or to  
119 Limnochordia (2,080) (Figure 2a and Supplementary Table 3). However, these might include  
120 innovations specific to these clades or linked to their phylogenetic placement and not necessarily to the  
121 diderm phenotype. Of the 15,964 families, 131 are shared uniquely between Halanaerobiales and  
122 Limnochordia, 73 between Limnochordia and Negativicutes, and 26 between Negativicutes and  
123 Halanaerobiales. Finally, 41 protein families are present in at least one member of each of the three  
124 diderm Firmicute lineages but are totally absent from monoderm Firmicutes, and will be referred to  
125 hereafter as the diderm “strict core” (Figure 2a, in bold). Consistently, these families include components  
126 of known OM systems (Supplementary Table 3) such as LpxABD (LPS synthesis), BamA (OM  
127 biogenesis) and FlgHI (Flagellar rings). However, some false negatives may arise from this approach.  
128 This is for example the case of LpxC and KdsABD (keto-deoxyoctulosonate (KDO) synthesis), which do  
129 not appear in the diderm strict core because homologues are present in a few monoderm genomes  
130 (Figure 3 and Supporting Data). In contrast, this analysis did not identify a strict core of monoderms,  
131 because, within the 3,863 protein families totally absent from diderm Firmicutes, only three are present  
132 in > 50% of monoderm genomes (Supplementary Table 3, highlighted in grey). Therefore, to relax our  
133 criteria, we implemented a complementary approach based on hierarchical clustering (HC) on the  
134 15,964 protein families (Methods). Among the 3,500 clusters generated by the HC approach, four (HC5,  
135 HC7, HC8 and HC9) appeared particularly enriched in the three diderm clades with respect to their  
136 monoderm relatives and contained at least one of the previously identified “strict core” diderm families  
137 present in more than 40% of diderm Firmicutes. These four clusters total 120 protein families  
138 (Supplementary Table 4), 95 of which were also retrieved with an alternative clustering approach  
139 (Methods, Supplementary Table 5). The HC5 (14 protein families), HC7 (67 protein families), and HC9  
140 (8 protein families) clusters show variable distribution patterns across diderm Firmicutes and appear to  
141 be mostly restricted to the Negativicutes (Figure 2b and Supplementary Table 4). In contrast, cluster  
142 HC8 has the sharpest pattern of enrichment in all three diderm Firmicutes clades (Figure 2b). It includes  
143 31 protein families, which comprise some of the OM markers that were missed by the strict  
144 presence/absence criterion, such as LpxC and KdsABD (Supplementary Table 4), confirming the

145 highest sensitivity of the clustering approach. Consistently with the strict core analysis, here again we  
146 could not highlight any cluster specifically enriched in monoderms with respect to diderms  
147 (Supplementary Data).

148 Finally, we used a third approach based on Pearson Correlation Coefficient (PCC) (Methods).  
149 We found that 83 families correlate with the diderm phenotype ( $PCC \geq 0.5$ ) (Figure 2c). Of these, 26  
150 are present in HC8, 56 are totally absent in monoderm Firmicutes, and 35 are present in all three diderm  
151 Firmicutes lineages (Supplementary Table 6). Some important OM markers were nevertheless not  
152 recovered by the analysis of protein families. This is for example the case of TamB, an inner membrane  
153 (IM) protein involved in the insertion of proteins in the OM. Another example is that of OmpM, a porin  
154 responsible for tethering the OM in Negativicutes, and some of the components of the Lpt system to  
155 export LPS to the OM (LptB and LptD). These homologues are not sufficiently conserved at the  
156 sequence level and ended up being split among different families. We therefore complemented the  
157 protein family approach by one based on protein domains (Methods). We identified all Pfam domains  
158 in our Firmicutes databank, and grouped proteins using three different approaches (Methods). We then  
159 computed the PCC of each of these groups with the diderm phenotype (Supplementary Figure 2 and  
160 Supplementary Table 7). This allowed to identify OM markers that were missed by the protein family  
161 approach, such as TamB and OmpM. All results (protein families and Pfam domains), were merged.  
162 For the proteins identified by more than one method, we kept the corresponding family or domain  
163 presenting the best PCC. Finally, we only kept the protein families or domains present in at least one  
164 member of the three diderm Firmicutes lineages. These analyses led to identify 52 protein  
165 families/domains that are specifically enriched in the three diderm Firmicutes lineages, the majority of  
166 which are either totally or mostly absent from monoderm Firmicutes and include crucial functions linked  
167 to the cell envelope and the OM (Figure 3a). Interestingly, they also include putative functions not  
168 immediately linked to OM biogenesis, such as RuvC (endonuclease) and RecX (a regulator of RecA,  
169 involved in DNA repair). Other proteins have only a generic annotation (e.g. Peptidase\_S55,  
170 Peptidase\_M23, AMIN) and might be linked to peptidoglycan (PG) related functions. Finally, some  
171 proteins are totally uncharacterized. For example, DUF3084 which is among the most highly correlated  
172 with the diderm phenotype, and is practically absent in monoderms. Finally, we identified 51 protein  
173 families and Pfam domains that correlate with the monoderm phenotype (Figure 3b, Supplementary  
174 Figure 2, Supplementary Table 8). However, and consistently with the strict core analysis, most of these  
175 proteins/domains are not widely distributed across monoderms, even those commonly assumed to be  
176 markers of Gram-positive (e.g. sortase, LPXTG anchor), and only one protein family (containing the S4  
177 domain, loosely annotated as involved in RNA binding and translation regulation) shows a PCC higher  
178 than 0.9. Therefore, unlike diderm Firmicutes, it appears that there is *de facto* no conserved core for  
179 monoderm Firmicutes.

180

181 *A large OM gene cluster is a distinguishing feature of all diderm Firmicutes and reveals potential novel functions related to*  
182 *OM biogenesis*

183 Very little experimental data are available on the nature of the cell envelope of diderm Firmicutes. In  
184 order to gather further information on the putative functions of the 52 family/domains most correlated  
185 with the diderm phenotype, we applied a “guilty by association” strategy by looking at their genomic  
186 environment. Twenty nine of them revealed to be part of a large gene cluster (up to 65 kb) that we  
187 previously identified in Negativicutes and Halanaerobiales<sup>10</sup>. This cluster, which will be hereafter  
188 referred to as the “OM cluster” is also present in all Limnochordia genomes with a very similar gene  
189 order (Figure 4) and therefore appears to be a distinguishing characteristic of all diderm Firmicutes.

190 The OM cluster codes for a number of important systems known to be involved in the biogenesis  
191 of the bacterial diderm cell envelope<sup>17,18</sup> (Figure 4). It contains the main players in the synthesis of LPS  
192 and its transfer to the OM (Figure 4, in blue). This includes the pathway responsible for synthesis of  
193 Lipid A, the innermost component of LPS (LpxACD/LpxI/LpxB/LpxK). While the enzymes for the  
194 core oligosaccharide component of LPS are present in Halanaerobiales and Negativicutes (WaaM,  
195 KdsABCD), they are absent altogether in Limnochordia, where they might be replaced by a large  
196 number of unrelated glycosyl transferases present in the cluster (Figure 4). All diderm Firmicutes seem  
197 to have a conserved LipidA flippase (MsbA), and the Lpt system for LPS transport to the OM  
198 (LptABC/LptFG/LptD)<sup>19</sup>. It is intriguing to note that the gene coding for LptD is not close to those  
199 encoding the other Lpt components, but frequently lies next to two genes with no annotated function or  
200 conserved domains in both Negativicutes and Halanaerobiales (Hypothetical 1 and 2 in Figure 4). This  
201 conserved three-gene arrangement suggests functional linkage and two potentially novel components of  
202 the machinery for LPS transport to the OM in diderm Firmicutes. Similarly, a gene coding for a third  
203 hypothetical protein is always present in the OM gene cluster (Figure 4, Hypothetical 3). No functional  
204 domains are annotated, but it’s strong conservation within the LPS genes strongly suggest some  
205 involvement in this pathway.

206 The OM cluster also includes another very important system responsible for the assembly of  
207 OM proteins (OMPs) into the lipid bilayer (Figure 4, in orange). It requires a coordinated process of  
208 folding into a  $\beta$ -barrel structure and membrane integration and it is accomplished by the  $\beta$ -barrel  
209 assembly machinery (BAM) complex whose main component is BamA (Omp85)<sup>17</sup>. In *Escherichia coli* and  
210 other Proteobacteria, an additional complex known as the translocation and assembly module (TAM) is  
211 present<sup>20</sup>. All three diderm Firmicutes lineages seem to possess a single system for OMP biogenesis  
212 composed of TamB and BamA, and one or multiple copies of the periplasmic chaperone Skp (Figure 4).  
213 This TamB/BamA complex was previously proposed to constitute an ancestral configuration predating  
214 the emergence of the Bam and Tam machineries in the evolution of Bacteria<sup>10,20</sup>. The TamB  
215 homologues of Limnochordia are longer than their counterparts in Halanaerobiales and Negativicutes  
216 and frequently have an N-terminal extension with no identifiable domains. This might indicate the  
217 existence of yet unknown novel functions or interactions of this machinery in diderm Firmicutes.

218 Curiously, the genome of *L. pilosa* and some other uncultured members of Limnochordia do not possess  
219 a homologue of the gene coding for TamB (Figure 4 and Supplementary Figure 3). It is unclear at this  
220 stage if these are true absences or repeated assembly errors.

221 One or multiple copies of the genes coding for OmpM can be observed in the OM cluster (Figure  
222 4, in red). OmpM is a porin with SLH domains that has been shown to attach non covalently the OM  
223 to a modified PG in Negativicutes, constituting an atypical OM tethering system that is different from  
224 the well-known Braun's lipoprotein<sup>11</sup>. This suggests that diderm Firmicutes may all use a similar strategy  
225 to attach their OM to PG. Curiously, in Halanaerobiales and Limnochordia, one of the *ompM* genes lies  
226 in a conserved context including homologues of genes coding for an ExbBD/TonB machinery,  
227 responsible for the active transport of molecules to the OM by exploiting the energized IM<sup>21</sup>. Moreover,  
228 these genes lie together with a gene coding for cohesin, a protein generally involved in DNA repair and  
229 gene regulation (Figure 4). Such strong synteny conservation may indicate a functional link among these  
230 proteins which is intriguing and will need experimental verification.

231 In Negativicutes only, the OM cluster contains a conserved four-gene arrangement coding for  
232 the IM components of the Mla machinery (MlaEFD) and the OM channel TolC (Figure 4, in green).  
233 The Mla system is responsible in diderm bacteria for maintaining lipid bilayer asymmetry and OM  
234 barrier by the transport of phospholipids from and to the IM<sup>22,24</sup>. No clear homologues of the periplasmic  
235 and OM components of the Mla system (MlaABC) could be identified in Negativicutes (Figure 4). This  
236 may suggest a novel mechanism to maintain lipid asymmetry in the Negativicutes involving TolC.  
237 Moreover, as Halanaerobiales and Limnochordia do not have any of these Mla coding gene homologues  
238 in their genomes, it remains to be understood how they cope with the absence of such crucial system for  
239 membrane integrity, or if they use a non-homologous machinery.

240 In flagellated diderm Firmicutes, the OM cluster also contains six genes encoding the flagellar  
241 proteins FlgFGAHIJ, including those for the specific P- and L-ring structures that in diderm bacteria  
242 serve to anchor the flagellum to the OM<sup>23</sup> (Figure 4, in pink).

243 Finally, a number of genes are highly conserved in the OM cluster in all three diderm Firmicutes  
244 clades but have only a generic annotation (Figure 4, in yellow). Some of these are included in a conserved  
245 six-genes arrangement (sometimes interrupted by flagellar genes) which codes for: a member of the S55  
246 peptidase family, a protein annotated as N-acetylglucosamine-1-phosphodiester alpha-N-  
247 acetylglucosaminidase (NAGPA) involved in protein glycosylation; a peptidase of the M23 family, a  
248 distant homologue of the lytic transglycosylase SpoIID, a homologue of the sporulation sigma factor  
249 SpoIIID, and a MreB homologue. Interestingly, a SpoIID homologue was recently shown to be involved  
250 in cell division in *Chlamydia trachomatis*<sup>24</sup> and it is tempting to speculate that some of these proteins are  
251 also involved in PG remodelling/daughter cell separation in diderm Firmicutes. Two other  
252 uncharacterized genes are always next to each other in the OM cluster (Figure 4): one codes for the  
253 uncharacterized protein DUF3084, and the other for RuvC (Holliday junction resolvase). Both proteins

254 are among most correlated with the diderm phenotype (see below, Figure 3), strongly suggesting  
255 functional interaction and a role in OM biogenesis.

256

257 *Phylogenomic analysis does not support acquisition of the OM by horizontal gene transfer*

258 The presence of the OM cluster may suggest that it was acquired by HGT. We previously showed by  
259 phylogenetic analysis of a concatenation of the four core LPS proteins (LpxABCD), that the sequences  
260 from Halanaerobiales and Negativicutes are closely related, match their reference species phylogeny,  
261 and do not stem from within another diderm bacterial phylum. We interpreted this result as support  
262 that these genes (and by extension the whole OM cluster) were not acquired via HGT, but were rather  
263 inherited from the ancestor of all Firmicutes, which would therefore have been a diderm with LPS<sup>5,10</sup>.  
264 Although in our opinion it is unlikely, the possibility of an acquisition in either the ancestor of  
265 Halanaerobiales or Negativicutes followed by a further HGT between the ancestors of these two clades  
266 remained open. We think that its presence in Limnochordia weakens this scenario, as this would imply  
267 an additional ancient transfer event. Nevertheless, in order to investigate further the HGT hypothesis,  
268 we searched for OM gene clusters similar to the one present in diderm Firmicutes in our Firmicutes  
269 databank as well as in 377 genomes representatives of major bacterial phyla (Methods). We could  
270 confirm that no other bacterial phylum possesses any gene cluster similar to the one in diderm Firmicutes  
271 (Supplementary Figure 3). In most diderm bacteria, the OM genes are in fact separated in a number of  
272 small clusters as in *E. coli*. Interestingly, bigger gene clusters could be observed in some diderm phyla  
273 that are evolutionarily close to the Firmicutes (Armatimonadetes and Synergistetes), but never as large  
274 as the one present in diderm Firmicutes (Supplementary Figure 3). Taken together, these results weaken  
275 the hypothesis of an acquisition of the OM in diderm Firmicutes.

276 An additional argument against the HGT scenario is that the OM cluster also contains genes  
277 that are not specific of diderm Firmicutes but are also present in monoderm lineages, such as *fabZ* (fatty  
278 acid metabolism), *murA* (cell wall synthesis), *mreB* (cell shape), *spoIID*, *spoIIID* (sporulation) (Figure 4).  
279 Many of these genes lie at the beginning of the OM cluster in diderm Firmicutes but are also similarly  
280 clustered in monoderm Firmicutes (Figure 4). Under the hypothesis of a diderm ancestor of all  
281 Firmicutes<sup>10</sup>, these genes could be remnants of an ancestral OM cluster which would have been lost in  
282 monoderms. Conversely, under the triple HGT scenario, it is difficult to explain why the OM cluster  
283 would have been inserted exactly at the same genomic position three times independently in  
284 Limnochordia, Halanaerobiales, and Negativicutes.

285 Finally, the HGT hypothesis is not supported by phylogenetic analysis. Among the OM markers  
286 encoded in the gene clusters previously calculated, we selected the ones most widely conserved in diderm  
287 bacteria. These were gathered into a large concatenated dataset, a strategy routinely used to increase  
288 ancient phylogenetic signal<sup>25</sup>. We built two alternative concatenations, one comprising 11 markers  
289 (FabZ, LpxACD/LpxB/LpxK, KdsBADC, LptB) totaling 146 taxa and 1,998 amino acid positions, and  
290 a second one also including the flagellar proteins (FlgFGAHIJ) totaling 3,705 amino acid positions but

291 less taxa (122 taxa) (Methods). A number of diderm phyla could not be included in the concatenation  
292 because they lacked more than half of these markers, or because their genomes did not have them in  
293 cluster (Supplementary Figure 6), preventing their clear identification. These datasets are larger than the  
294 LPS concatenation we analysed previously<sup>10</sup>. Given the small number of positions with respect to the  
295 large evolutionary distances analysed, ML trees from both concatenations are not totally resolved, but  
296 display nevertheless a topology in overall agreement with known taxonomy (Figure 5a and  
297 Supplementary Figure 4, respectively). In particular, we observe a well-supported monophyly of major  
298 bacterial diderm phyla, and an overall topology that is consistent with the known relationships within  
299 Bacteria, notably the two large clades of Terrabacteria and Gracilicutes<sup>26,27</sup>. These results indicate that  
300 the OM markers were present in the ancestors of each of these major diderm phyla and have not been  
301 subjected to extensive HGT during bacterial diversification. Consistently, we observe the monophyly of  
302 the three diderm Firmicutes clades, which indicates that their OM markers are more closely related  
303 among them than to other bacteria, and do not stem from within any specific diderm bacterial phylum,  
304 which would have been expected if these markers were acquired from HGT (Figure 5a). Moreover, that  
305 clade formed by diderm Firmicutes groups with other Terrabacteria phyla, in agreement with the  
306 phylogenetic placement of the Firmicutes.

307 Together, these results indicate that an OM with LPS was already present in the ancestor of the  
308 Halanaerobiales, Negativicutes, and Limnochordia, which is by definition the ancestor of all Firmicutes.  
309 They strengthen the diderm-first hypothesis for this phylum<sup>5</sup>, and the emergence of the monoderm cell  
310 envelope by multiple losses of the OM (Figure 5b). Moreover, they support the fact that LPS-OM have  
311 a very ancient origin in Bacteria and were inherited remarkably vertically throughout the diversification  
312 of the major diderm phyla.

313

## 314 **Discussion**

315 Past hypotheses have supported either a monoderm-first<sup>1,3,4,28</sup> or a diderm-first scenario<sup>2</sup> for Bacteria<sup>5</sup>.  
316 However, most of them did not consider the phylogenetic relationships among diderm and monoderm  
317 lineages and among outer envelope markers. The existence of three independent diderm lineages within  
318 the Firmicutes adds an important piece to the puzzle and shows that, at least in this phylum, the OM is  
319 ancestral and the monoderm phenotype is a derived character which arose multiple times independently  
320 through OM loss. The hypothesis of an ancestor with an LPS-OM is also supported by the evidence that  
321 the phylogeny of the Firmicutes is becoming increasingly populated by diderm lineages, notably at its  
322 deepest offshoots, and it cannot be excluded that even more will be discovered in the future.

323 How the OM would have been lost multiple times in the Firmicutes remains to be understood.  
324 Endosporulation was probably important in the transition between monoderms and diderms<sup>3,29</sup>. In fact,  
325 during the process of endosporulation the cell produces a spore that is transiently surrounded by a second  
326 membrane, which is then lost during maturation in sporulating monoderm Firmicutes, while is retained  
327 in sporulating diderm Firmicutes<sup>30</sup>. While previous hypotheses proposed that endosporulation would

328 have allowed the emergence of the OM<sup>3,4,29,30</sup>, we rather think that the opposite occurred, i.e. that viable  
329 accidents during endospore formation allowed multiple OM losses in the Firmicutes<sup>5</sup> (Figure 5b). The  
330 widespread presence and antiquity of endospore formation in this phylum (Supplementary Figure 1,  
331 Methods) could have allowed multiple occurrences of such accidents during its diversification, and this  
332 is probably the reason why the Firmicutes are currently the only phylum containing both monoderm  
333 and diderm envelopes (the presence of OMs in some Actinobacteria is likely an independent  
334 convergence, see below). The alternative scenario where the OM would have been acquired via multiple  
335 HGT events in the Firmicutes remains possible, but we believe it is less supported by our data. Moreover,  
336 rather than suggesting HGT, the clustering of the OM genes in diderm Firmicutes may indicate a tight  
337 coordination of the various OM biogenesis processes, which could represent a peculiarity of these  
338 lineages. Finally, if the OM was acquired multiple times by HGT, this is not simpler as a process to  
339 imagine, as it would have necessitated that all the complex machineries for OM biogenesis become  
340 immediately functional in a monoderm context, notably attaching the OM to the PG wall and stabilizing  
341 it, adapting existing flagella and secretion systems to span two membranes, and developing transport of  
342 key compounds to the OM.

343         It may be argued that the benefit of having an OM is such that there would have been no  
344 selective advantage in losing it. However, no advantage has to be invoked if the loss of the OM was the  
345 result of viable accidents making it unstable. These might have not led to a decrease in fitness so dramatic  
346 as to immediately disadvantage the resulting monoderm phenotype, in particular if accompanied (either  
347 preceded or followed) by an increase in thickness of the cell wall<sup>2,5,10</sup>. Moreover, the monoderm  
348 phenotype with a thick PG wall might have resulted advantageous in some specific environmental  
349 conditions (e.g. drought, high temperature). The absence of a core of protein families specific to  
350 monoderm Firmicutes may reflect the fact that different solutions were found independently to  
351 accommodate each of these multiple OM loss events, and it is interesting to note that PG structure is  
352 indeed highly variable in the Firmicutes<sup>29</sup>, and likely also the arsenal of enzyme families needed to  
353 produce it and remodel it. Following loss of the OM, the genes involved in its biogenesis would have  
354 been progressively lost, and some perhaps repurposed for cell envelope functions in monoderm  
355 Firmicutes, a hypothesis that will need specific analysis and experimental evidence. The availability of  
356 genetic tools in the Negativicute *Veillonella parvula*<sup>5</sup> opens the way to test experimentally some of these  
357 hypotheses and will allow to gather precious insights about the biogenesis and functioning both the  
358 diderm and the monoderm cell envelope.

359         Whether the last bacterial common ancestor (LBCA) also had an OM has been unclear, notably  
360 due to uncertainties in the phylogenetic relationships among diderms and monoderms. We calculated a  
361 phylogeny including all the main bacterial phyla used in this study, and we inferred their diderm or  
362 monoderm nature of their cell envelope by mapping the presence or absence, respectively, of two key  
363 markers of the OM, BamA and LpxA (Figure 6a, Supplementary Table 9, Methods). Although the cell  
364 envelope remains uncharacterized in most phyla, notably the uncultured one for which the diderm or

365 monoderm status can only be tentatively inferred in-silico<sup>5,31</sup>, it is already clear that the presence of  
366 diderms is overwhelming in Bacteria as compared to monoderms. Whether the LBCA was already a  
367 diderm or a monoderm will require to firmly establish the root of Bacteria, a complex methodological  
368 issue that can only be solved by specific analyses. We chose here to display a root in between  
369 Terrabacteria and Gracilicutes, supported by our recent phylogenomic analysis<sup>32</sup>. Combined with our  
370 evidence that the LPS-OM did not spread across diderm bacteria by HGT but was rather inherited  
371 vertically (Figure 5a), this root may imply that the LBCA could have been a diderm (Figure 6b, top).  
372 Alternative roots have been proposed in the literature, such as a possible one lying in the monoderm  
373 Chloroflexi or in the Candidate Phyla Radiation<sup>3,33,34</sup>, which could support an LBCA with one  
374 membrane (Figure 6b, bottom). Most importantly, and irrespective of whether the LBCA was a diderm  
375 or not, our results show that monoderm phyla do not constitute a natural group but are polyphyletic.  
376 This indicates that the monoderm cell envelope would in any case have appeared multiple times  
377 independently through OM loss, and working out all the evolutionary paths that led to these transitions  
378 is in our opinion the most important and challenging goal.

379         It remains unclear what was the mechanism involved in the loss of the OM in the monoderm  
380 phyla other than the Firmicutes, but it may have involved different types of viable accidents causing an  
381 instability of the OM. It is evident that the few known monoderm phyla are only present in the large  
382 clade of Terrabacteria which displays a larger range of cell envelopes with respect to the more  
383 homogenous Gracilicutes that are composed only of diderms mostly endowed with LPS (Figure 6a). This  
384 larger diversity possibly suggests that the Terrabacteria cell envelopes have retained some ancestral  
385 characteristics. As such, diderm Firmicutes could become good experimental models for the primordial  
386 bacterial cell envelope.

387         Currently, only the Firmicutes and the Actinobacteria display a mixture of monoderm and  
388 diderm lineages. However, the presence of the mycolic acid OM in Actinobacteria such as  
389 *Corynebacteria*<sup>35</sup> is likely an independent *de novo* origination, as these taxa do not possess any of the  
390 classical OM markers. The presence of complex cell envelope structure and potential OM-like structures  
391 has also been recently reported in the Chloroflexi<sup>36</sup>, a phylum that lacks all classical OM markers and  
392 thought to be monoderm<sup>37</sup>. However, virtually nothing is known on the cell envelope structure of most  
393 bacterial phyla, in particular those with no representative cultured members, and obtaining  
394 ultrastructural data for these phyla is an important challenge of future research.

395         Finally, how the OM initially came into being remains unknown. It is possible that it arose from  
396 a simpler cell surrounded by a single membrane (e.g. monoderm type), but we have no means by using  
397 phylogeny to go this far back in time beyond the LBCA. However, we think it unlikely that  
398 endospore formation was at the origin of the OM in the LBCA<sup>3,4</sup>, as today this type of sporulation that  
399 produces a transient OM is specific of the Firmicutes and likely originated in this phylum. Alternatively,  
400 early speculation by Blobel postulated that the very first cell was already surrounded by a double  
401 membrane through a mechanism involving the formation of a “gastruloid” vesicle and the fusion of its



402 extremities<sup>38</sup>. So, it is possible that early life never went through a “simpler” monoderm phase but started  
403 already as diderm.

404 Some of these questions should be addressed in the future through a more systematic  
405 characterization of a wide range of cell envelopes, both monoderm and diderm, across all bacterial phyla  
406 -notably the uncultured ones-, combined with large-scale comparative genomics and phylogenetic  
407 analysis to fully reconstruct the evolutionary history that accompanied their evolution and the multiple  
408 transitions among them, as well as the development of experimental models from unexplored branches  
409 of the Tree of Life.

410

#### 411 **Acknowledgments**

412 S.G., C.B. and J.W. wish to acknowledge funding from the French National Research Agency (ANR),  
413 project Fir-OM (ANR-16-CE12-0010) and from the Institut Pasteur Programmes Transversaux de  
414 Recherche (PTR 39-16). D.M. and D.P were supported by the Pasteur-Paris University (PPU)  
415 International PhD Program. This work used the computational and storage services (TARS cluster)  
416 provided by the IT department at Institut Pasteur, Paris.

417

#### 418 **Author Contributions**

419 S.G. conceived the study. N.T. and D.M. carried out all comparative genomics and phylogeny analyses.  
420 J. W. helped with annotation of the OM markers. D.P and G.B. helped with genome reconstruction of  
421 two uncultured Limnochordia genomes in an earlier version of the study. P. A. assembled the DB  
422 Bacteria and calculated the reference tree shown in Figure 6. C.B. helped with annotation and overall  
423 supervision. N.T., C.B. and S.G. wrote the paper, with contribution from D.M and J.W. All authors  
424 have read and approved the manuscript.

425

#### 426 **Material and Methods**

427

##### 428 *Updating the Firmicutes reference tree and identification of diderm lineages*

429 We retrieved 1,639 genomes annotated as Firmicutes from the dataset of Parks et al.<sup>13</sup>, deposited under  
430 NCBI BioProject PRJNA348753. These genomes were isolated from different environments and their  
431 genomes quality goes from partial to near complete (Supplementary Table 1). According to the  
432 taxonomy provided by Parks et al.<sup>13</sup>, these genomes consist in 351 Bacilli, 980 Clostridia, 61  
433 Erysipelotrichia, 1 Tissierellia, 62 Negativicutes, 1 Halanaerobiales and 183 annotated as unclassified  
434 Firmicutes. Because these UBA genomes were in the nucleotide format at the time of this analysis, we  
435 used Prodigal<sup>39</sup> with default parameters to predict genes. To analyze their phylogenetic placement within  
436 the reference phylogeny of Firmicutes, we added 230 complete genomes from representative of all  
437 families available in the NCBI databases as for December 2017, including 61 Bacilli, 86 Clostridia, 4  
438 Tissierellia, 62 Negativicutes, 16 Halanaerobiales and the genome of the only available representative

439 of Limnochordia, *Limnochorda pilosa*. This resulted in a databank of 1,869 genomes (DB LARGE  
440 Firmicutes) (Supporting Data).

441 Exhaustive Hidden Markov Model (HMM)-based homology searches (with an e-value cutoff of 1e-04)  
442 were carried out by using HMM profiles of the complete set of 54 bacterial ribosomal proteins from the  
443 Pfam 29.0 database<sup>40</sup> as queries using the HMMER package<sup>41</sup>. Absences were checked with  
444 TBLASTN<sup>42</sup>. 45 ribosomal proteins present in > 70% of the genomes were kept for analysis. 514 UBA  
445 Firmicutes genomes having less than 35 ribosomal proteins were considered as too partial and discarded  
446 from analysis. 13 taxa were included as outgroup (Supporting Data). The 45 ribosomal proteins of the  
447 1,355 remaining UBA and outgroup genomes were aligned by CLUSTAL OMEGA<sup>43</sup> with default  
448 parameters and trimmed using BMGE-1.1<sup>44</sup> with the BLOSUM35 substitution matrix. The resulting  
449 trimmed alignments were concatenated into a supermatrix (1,368 taxa and 5,087 amino acid positions).  
450 A ML tree was generated using IQ-TREE v1.4.4<sup>45</sup>, with the ultrafast bootstrap approximation<sup>46</sup>  
451 imposed by the very large size of the dataset and the C60-profile mixture model LG+C60+F+G, which  
452 is a variant of the CAT model<sup>47</sup> for ML analysis.

453 To identify new diderm lineages among the UBA genomes, we used HMMSEARCH with a cutoff e-  
454 value of 1e-04 to screen them for the proteins involved in the first conserved steps of LPS synthesis  
455 (LpxABCD) (TIGR01852, PF02684, PF03331, PF04613), and for other protein domains previously  
456 used as markers of Gram-negativity<sup>7</sup>: Omp85 (PF01103), POTRA (PF07244), ExbD (PF02472), Secretin  
457 (PF00263), TamB (PF04357) and TonB\_C (PF03544).

458 In order to identify sporulating taxa, we used HMMSEARCH to screen the DB SMALL Firmicutes  
459 using the Pfam domain spo0A\_C (PF08769) with the option -cut\_ga, and we mapped the results onto  
460 the corresponding tree of Firmicutes (Supporting Data).

461 All trees were annotated using iTol<sup>48</sup>.

462

#### 463 *Distribution of protein families and domains in diderm and monoderm Firmicutes*

464 To carry out the large-scale comparative genomic analysis, we assembled a reduced databank of 316  
465 genomes. It includes the 230 reference Firmicutes genomes and the newly identified UBA diderm  
466 Firmicutes (46 Limnochordia, 39 Negativicutes, 1 Halanaerobiales), for a total of 165 diderm and 151  
467 monoderm taxa (DB SMALL Firmicutes) (Supporting Data). The 861,409 proteins contained in the DB  
468 SMALL Firmicutes were annotated by using eggNOG-Mapper<sup>49</sup> with default parameters. The  
469 eggNOG-Mapper tool uses precomputed orthologous groups and phylogenies from the eggNOG  
470 database<sup>50</sup> to transfer functional information from fine-grained orthologs only. In a second approach,  
471 Pfam domains were predicted for each protein using HMMSEARCH (e-value <= 1e-5) against the  
472 Pfam 29.0 database. The results of the two approaches were merged and each protein family was  
473 annotated according to the most frequent prediction of its members.

474 We performed all vs all pairwise comparisons of protein sequences contained in the DB SMALL  
475 Firmicutes using BLASTP v2.6.0 with default parameters. Protein families were assembled with SILIX

476 v1.2.9<sup>51</sup>. Identity thresholds values from 30% to 60% with intervals of 5% were tested, with a coverage  
477 of at least 80%. The resulting protein families were then refined using HIFIX v1.0.5, which performs a  
478 three-step high-quality sequence clustering guided by network topology and multiple alignment  
479 likelihood<sup>52</sup>. To assess the most suitable identity threshold to group orthologous proteins, we tested  
480 different cutoffs by using as positive control the clustering of 16 ribosomal proteins commonly used in  
481 phylogenetic analyses and of the four core LPS proteins (LpxABCD). The identity threshold that  
482 maximized the number of true positives and minimized the number of false positives was 35%. This  
483 80% coverage-35% identity cutoff cannot however completely exclude some false negatives or false  
484 positives. Applying this threshold resulted in 176,024 protein families.

485 From these, we retained the families present in at least five taxa, resulting in 15,964 families for further  
486 analysis (Supporting Data), and a presence/absence matrix was built. Among the 15,964 families, 41  
487 were completely absent from monoderm Firmicutes while present in at least one member of all three  
488 diderm lineages (“strict core diderm families”).

489 In order to relax this strict criterium we used two clustering approaches on the presence/absence matrix  
490 (HCLUST and K-MEANS, both implemented in R). HCLUST clusters hierarchically the families  
491 according to Jaccard distances calculated on the presence/absence matrix. We tested different cutoffs  
492 and chose the one that allowed gathering the four core LPS protein families (LpxABCD) in the same  
493 cluster as a positive control (number of generated clusters set to 3,500) (Supporting Data). Four clusters,  
494 HC5, HC7, HC8 and HC9 included at least one of the “strict core” diderm families with a large  
495 taxonomic distribution (present in more than 40% of the diderm Firmicutes genomes), totalling 120  
496 families.

497 For the second method, we clustered families using K-MEANS (k = 500) based on Multiple  
498 Correspondence Analysis (MCA). As the K-MEANS approach involves defining a random set of starting  
499 points in a multidimensional space, 10 iterations were run. 173 protein families clustered with at least  
500 one of the 41 strict core diderm families and were present in more than 40% of the diderm Firmicutes  
501 genomes in all K-MEANS iterations. Among these families, 95 were common to HCLUST and K-  
502 MEANS.

503 In parallel to the distribution of protein families, we used a second approach based on protein domains.  
504 Because proteins can contain different domains or multiple occurrences of the same domain, we used  
505 three different approaches: in the first, we considered together all proteins containing exactly the same  
506 type and number of predicted domains (ALL); in the second, we considered together proteins containing  
507 the same domains even if some had more than one occurrence (COLLAPSED); in the third, we counted  
508 the same proteins as many times as the different domains they contain (SINGLE) (Supplementary Tables  
509 7 and 8) (Supporting Data).

510 Finally, the four obtained datasets (protein families, and the three Pfam domain approaches) were used  
511 to identify the diderm and monoderm specific protein families and Pfam domains using the Pearson  
512 correlation coefficient (PCC) based on their presence/absence in the diderm and monoderm taxa (higher

513 than 0.5). The conserved genomic locus for cell envelope components in the Firmicutes was assessed  
514 using GeneSpy<sup>53</sup>.

515

#### 516 *Distribution of the OM cluster in Bacteria and evolutionary analysis*

517 In order to study the distribution of the OM cluster, we built HMM profiles of all genes included in the  
518 OM cluster of diderm Firmicutes. Then, we used MacSyFinder<sup>54</sup> to identify clusters containing these  
519 genes in the DB SMALL Firmicutes and a second databank including 377 genomes representative of 58  
520 main bacterial phyla (DB Bacteria) (Supporting Data). We defined a cluster as a system with at least  
521 three of these OM components with a separation no greater than five other genes.

522 Among the gene clusters identified above, we selected 11 OM markers most conserved across Bacteria  
523 (FabZ, KdsA, KdsB, KdsC, KdsD, LptB, LpxA, LpxB, LpxC, LpxD, LpxK\_WaaA). For each of them,  
524 homologues were aligned with MAFFT using the L-INS-I option<sup>55</sup> and trimmed using BMGE-1.1. The  
525 resulting alignments were concatenated by allowing a maximum of six missing markers per taxon and  
526 leading to a supermatrix of 146 taxa and 1,998 amino acid position. A second matrix including the  
527 flagellar components FlgF, FlgG, FlgA, FlgH, FlgI and FlgJ was assembled by allowing a maximum of  
528 eight missing markers per taxon and included 122 taxa (because many do not have flagella) and 3,705  
529 amino acid positions.

530 For the reference phylogeny of Bacteria, we used a concatenation of RNA polymerase subunits B, B',  
531 and IF-2 (2,144 amino acid positions and 377 taxa).

532 For all concatenations, ML trees were generated using IQ-TREE v1.4.4 with the profile mixture model  
533 LG+C60+F+G with ultrafast bootstrap supports calculated on 1,000 replicates from the original data.

534 In order to map cell envelope types onto the reference phylogeny Bacteria, we built an HMM profile for  
535 BamA using the sequences in the corresponding family described above, and used it together with the  
536 HMM profile of LpxA to screen the DB Bacteria with HMMSEARCH (e-value  $\leq 1e-4$ ). Results were  
537 then refined and absences checked with TBLASTN. All trees were annotated using iTOL<sup>56</sup>.

538

#### 539 **Declaration of Interests**

540 The authors declare no competing interests.

541

#### 542 **Data availability**

543 All raw data relative to this analysis (databanks, sequence accession numbers, sequence datasets and  
544 corresponding trees, protein families) are provided as Supporting Data.

545 <https://data.mendeley.com/datasets/3pcn9779gc/draft?a=8cc3c448-b3e7-4b02-96fb-9b3a0af4625d>

546

#### 547 **Figure legends**

548 **Figure 1: An updated reference phylogeny of the Firmicutes reveals a third diderm clade.**

549 ML reference tree of the Firmicutes including 1,125 UBA Firmicutes genomes based on concatenation  
550 of 45 ribosomal proteins (1,368 taxa, 5,087 amino acid positions). The tree was inferred with IQ-TREE  
551 1.4.4 using the LG+C60+F+G model. Node supports higher than 95% are displayed. The tree is rooted  
552 with representatives of major bacterial taxa. The scale bar corresponds to the average number of  
553 substitutions per site. Names in red correspond to groups containing UBA genomes only and no  
554 representative of any known family as for December 2017. Coloured triangles indicate the three diderm  
555 clades. For the corresponding full phylogeny see Supplementary Data.

556

557 **Figure 2: Phyletic patterns of protein families highlight the functional core of the diderm**  
558 **Firmicutes OM.**

559 (a) Venn diagram showing the distribution of 15,964 protein families of the Firmicutes pan-proteome in  
560 Negativicutes, Halanaerobiales, Limnochordia and monoderm Firmicutes.

561 (b) Excerpt of the HC-based distribution of the 15,964 protein families of the Firmicutes pan-proteome  
562 with a focus on the four HC clusters (HC5, HC7, HC8 and HC9) containing the 120 protein families  
563 composing the relaxed core (columns in red). (c) Distribution of the 15,964 protein families of the  
564 Firmicutes pan-proteome. Dots in purple and pink correspond to protein families with a PCC > 0.5 with  
565 the diderm or monoderm phenotype, respectively.

566

567 **Figure 3: Distribution and functional annotation of the protein families and PFAM**  
568 **domains highly correlated with the diderm phenotype (a) and the monoderm phenotype**  
569 **(b).**

570 (a) 52 protein families and Pfam domains correlated with the diderm phenotype (PCC > 0.5) and  
571 present in at least one member of each of the three diderm lineages. When a protein family and a  
572 Pfam domain corresponded, we chose to display the one with the best correlation. Presence in the OM  
573 gene cluster is also indicated. Superscripts indicate proteins which correspond to two different domains  
574 (TamB and OmpM).

575 (b) 51 protein families and Pfam correlated with the monoderm phenotype. When a protein family and  
576 a Pfam domain corresponded, the one with the best correlation is displayed.

577

578 **Figure 4: A large OM gene cluster is a distinguished feature of all diderm Firmicutes.**

579 Annotation of the genomic locus containing OM-related markers in all three diderm Firmicute lineages.  
580 Color codes: blue (LPS synthesis and transfer); orange (OMP biogenesis); red (OM tethering); green  
581 (Lipid asymmetry); pink (Flagellum); yellow (unclear function); white (no functional domains detected);  
582 gray (no direct link to OM); hashed: three conserved hypothetical proteins possibly related to the OM.  
583 Lateral bars mean separate loci in the genome. Vertical bars mean end of a contig in unclosed genomes.  
584 The organization of these genes in *E. coli* is given as a comparison. Monoderm Firmicutes (Clostridiales

585 and Bacilli) are also displayed to illustrate the presence of the beginning of the OM cluster, a possible  
586 remnant of a diderm past. All accession numbers are provided in Supplementary Data.

587

588 **Figure 5: Phylogenomic analysis does not support acquisition of the OM by horizontal**  
589 **gene transfer (a) and supports multiple and independent losses of the OM (b).**

590 (a) ML tree based on the concatenation of 11 conserved OM markers (FabZ, KdsA, KdsB, KdsC, KdsD,  
591 LptB, LpxA, LpxB, LpxC, LpxD, LpxK\_WaaA) including 14 diderm phyla and 1,998 amino acid  
592 positions. The tree was inferred with IQ-TREE 1.4.4 and the LG+C60+F+G model. Node supports  
593 higher than 70% are displayed. The scale bar corresponds to the average number of substitutions per  
594 site. In the absence of an outgroup, the tree is tentatively rooted in between Terrabacteria and  
595 Gracilicutes<sup>32</sup>. Even though the tree only contains diderm phyla, it shows that the presence of main OM  
596 markers predates the divergence of these phyla, including that of Firmicutes and that therefore an OM  
597 with LPS has an ancient origin. For the corresponding full phylogeny see Supplementary Data.

598 (b) Evolutionary scenario for the origin and evolution of the OM in the Firmicutes mapped on a  
599 schematic of the reference phylogeny in Figure 1. The ancestor of Firmicutes is indicated as a sporulating  
600 diderm with LPS. The LPS-OM was inherited in the three diderm lineages (Halanaerobiales,  
601 Limnochordia and Negativicutes), while it was lost three times independently to give rise to the classical  
602 monoderm cell envelope.

603

604 **Figure 6: Distribution of monoderm and diderm cell envelopes across Bacteria and two**  
605 **potential evolutionary scenarios for their origin.**

606 (a) ML reference tree of Bacteria, with cell envelope types mapped on it. The tree was obtained from a  
607 concatenation of RNA polymerase subunits B, B', and translation initiation factor IF-2 (2,144 amino  
608 acids positions and 377 taxa). The tree was inferred with IQTREE and the LG+C60+G+F model.  
609 Node supports higher than 70% are displayed. The scale bar represents the average number of  
610 substitutions per site. The tree does not include an outgroup but is tentatively rooted between two large  
611 clades corresponding to Terrabacteria and Gracilicutes, as in Raymann et al.,<sup>32</sup>. For the corresponding  
612 full phylogeny see Supplementary Data. PVC s.l. (*sensu latu*) indicates the clade including the PVC  
613 superphylum (Planctomycetes, Verrucomicrobia, Chlamydiae) and relative phyla; FCB s.l. (*sensu latu*)  
614 indicates the clade including the FBC superphylum (Fibrobacteres, Bacteroidetes, Chlorobi) and relative  
615 phyla; Proteobacteria s.l. (*sensu latu*) indicates the clade including the Proteobacteria subdivisions and  
616 related phyla. For the corresponding full phylogeny see Supplementary Data.

617 (b) Evolutionary scenario mapped on a schematic version of the tree in (a). Different roots of the Bacteria  
618 have been proposed in the literature, we show here two alternative ones as an example: (i) in between  
619 Terrabacteria and Gracilicutes<sup>32</sup> (top) or (ii) in the branch leading to the CPR and Chloroflexi<sup>33,34</sup>  
620 (bottom). In the first scenario, the LBCA could have been a diderm. In the second scenario, the LBCA  
621 could have been a monoderm and the OM would have appeared just after the divergence of Chloroflexi

622 and CPR. Note that both scenarios imply multiple losses of the OM. The second scenario also leaves  
623 open the possibility that the LBCA was a diderm and that the OM was lost in the branch leading to the  
624 Chloroflexi and the CPR.

625

## 626 **References**

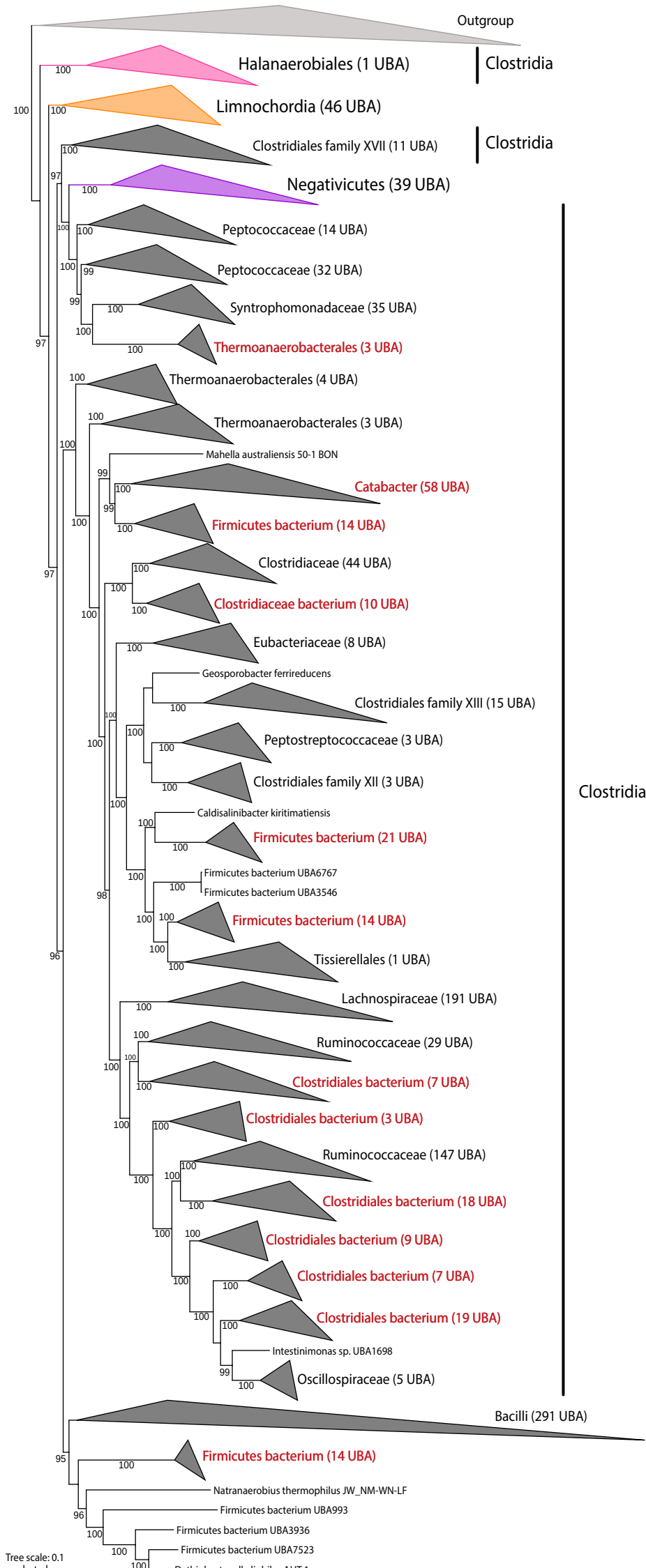
- 627 1. Gupta, R. S. Origin of diderm (Gram-negative) bacteria: Antibiotic selection pressure rather  
628 than endosymbiosis likely led to the evolution of bacterial cells with two membranes. *Antonie van*  
629 *Leeuwenhoek, International Journal of General and Molecular Microbiology* vol. 100 171–182 (2011).
- 630 2. Cavalier-Smith, T. The neomuran origin of archaeobacteria, the negibacterial root of the  
631 universal tree and bacterial megaclassification. *Int. J. Syst. Evol. Microbiol.* **52**, 7–76 (2002).
- 632 3. Tocheva, E. I., Ortega, D. R. & Jensen, G. J. Sporulation, bacterial cell envelopes and the  
633 origin of life. *Nat. Rev. Microbiol.* **14**, 535–542 (2016).
- 634 4. Errington, J. L-form bacteria, cell walls and the origins of life. *Open Biol.* **3**, 120143 (2013).
- 635 5. Megrian, D., Taib, N., Witwinowski, J., Beloin, C. & Gribaldo, S. One or two membranes?  
636 Diderm Firmicutes challenge the Gram-positive/Gram-negative divide. *Molecular Microbiology*  
637 vol. 113 659–671 (2020).
- 638 6. Mavromatis, K. *et al.* Genome analysis of the anaerobic thermohalophilic bacterium  
639 *Halothermothrix orenii*. *PLoS One* **4**, (2009).
- 640 7. Tocheva, E. I. *et al.* Peptidoglycan Remodeling and Conversion of an Inner Membrane into an  
641 Outer Membrane During Sporulation *Elitza*. *Cell* **146**, 799–812 (2012).
- 642 8. Campbell, C., Sutcliffe, I. C. & Gupta, R. S. Comparative proteome analysis of  
643 *Acidaminococcus intestini* supports a relationship between outer membrane biogenesis in  
644 Negativicutes and Proteobacteria. *Arch. Microbiol.* **196**, 307–310 (2014).
- 645 9. Helander, I. M., Hurme, R., Haikara, A. & Moran, A. P. Separation and characterization of  
646 two chemically distinct lipopolysaccharides in two *Pectinatus* species. *J. Bacteriol.* **174**, 3348–  
647 3354 (1992).
- 648 10. Antunes, L. C. *et al.* Phylogenomic analysis supports the ancestral presence of LPS-outer  
649 membranes in the firmicutes. *Elife* **5**, e14589 (2016).
- 650 11. Kojima, S. *et al.* Cadaverine covalently linked to peptidoglycan is required for interaction  
651 between the peptidoglycan and the periplasm-exposed S-layer-homologous domain of major  
652 outer membrane protein Mep45 in *Selenomonas ruminantium*. *J. Bacteriol.* **192**, 5953–5961  
653 (2010).
- 654 12. Poppleton, D. I. *et al.* Outer membrane proteome of *Veillonella parvula*: A diderm firmicute of  
655 the human microbiome. *Front. Microbiol.* **8**, 1–17 (2017).

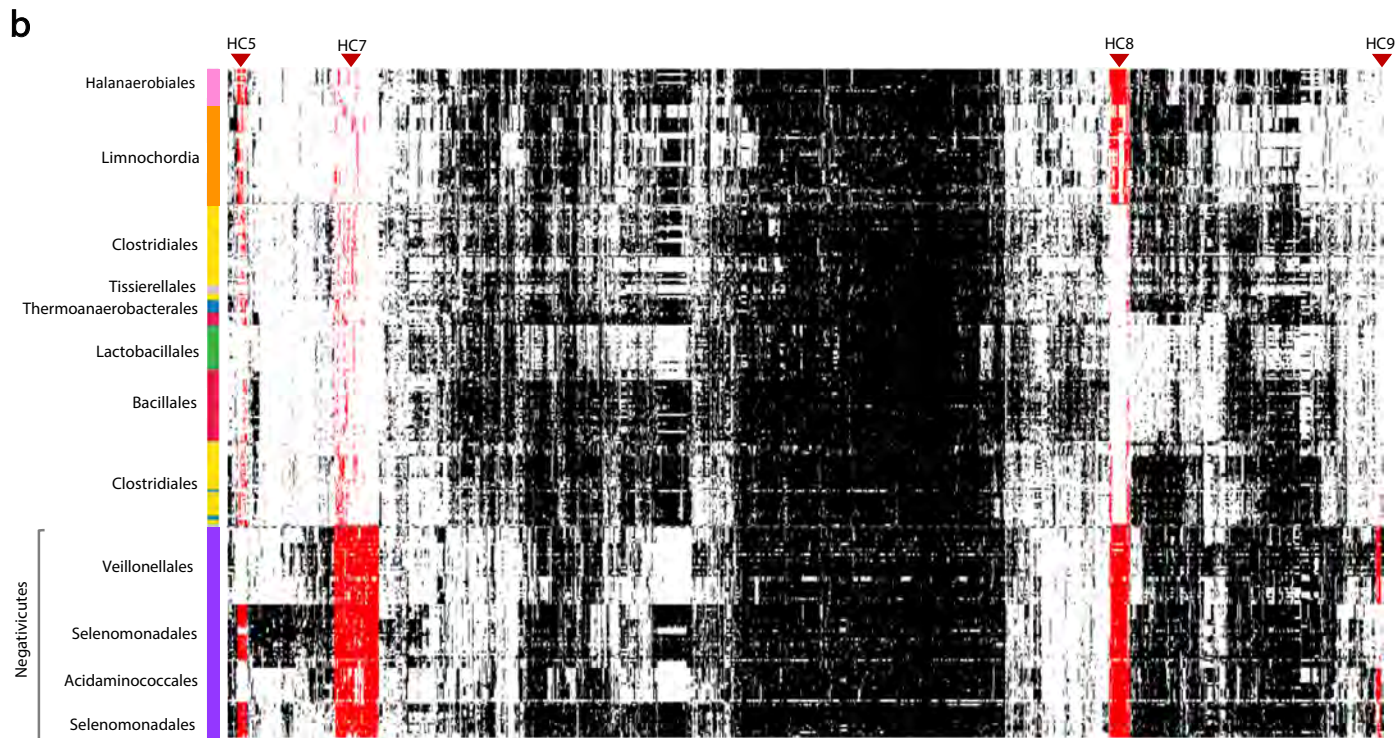
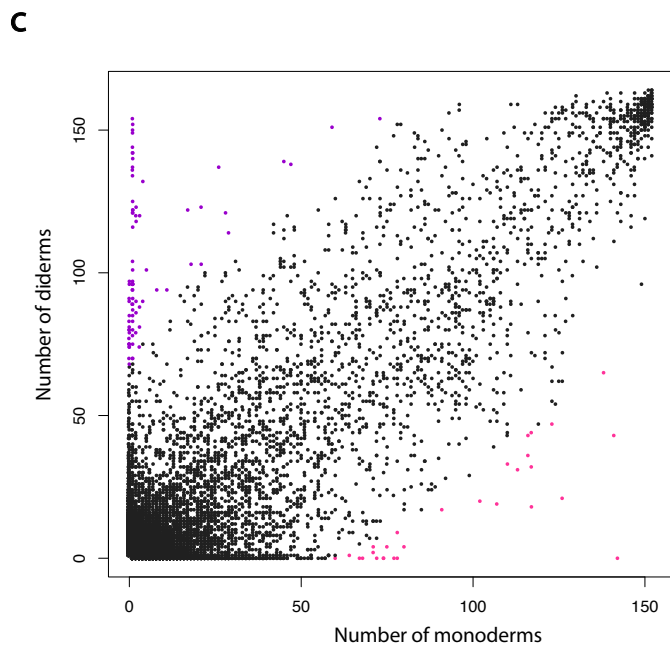
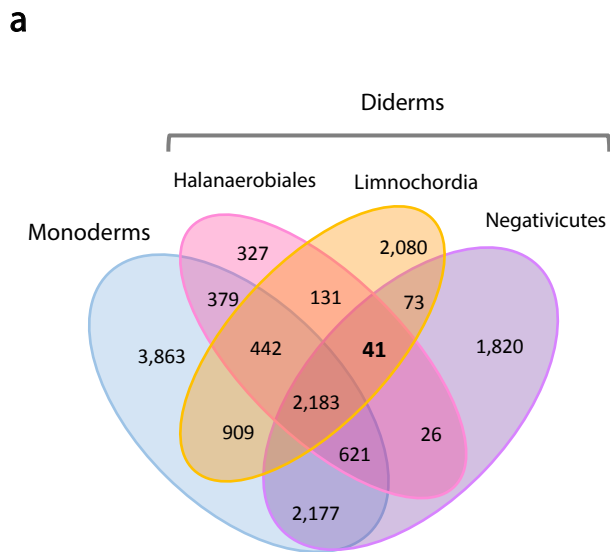
- 656 13. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially  
657 expands the tree of life. *Nat. Microbiol.* **2**, (2017).
- 658 14. Yutin, N. & Galperin, M. Y. A genomic update on clostridial phylogeny: Gram-negative spore  
659 formers and other misplaced clostridia. *Environ. Microbiol.* **15**, 2631–2641 (2013).
- 660 15. Watanabe, M., Kojima, H. & Fukui, M. *Limnochorda pilosa* gen. nov., sp. nov., a moderately  
661 thermophilic, facultatively anaerobic, pleomorphic bacterium and proposal of  
662 *Limnochordaceae* fam. nov., *Limnochordales* ord. nov. and *Limnochordia* classis nov. in the  
663 phylum Firmicutes. *Int. J. Syst. Evol. Microbiol.* **65**, 2378–2384 (2015).
- 664 16. Watanabe, M., Kojima, H. & Fukui, M. Complete genome sequence and cell structure of  
665 *Limnochorda pilosa*, a Gram-negative spore-former within the phylum Firmicutes. *Int. J. Syst.*  
666 *Evol. Microbiol.* **66**, 1330–1339 (2016).
- 667 17. Silhavy, T. J., Kahne, D. & Walker, S. The bacterial cell envelope. *Cold Spring Harbor perspectives*  
668 *in biology* vol. 2 a000414–a000414 (2010).
- 669 18. Bos, M. P., Robert, V. & Tommassen, J. Biogenesis of the gram-negative bacterial outer  
670 membrane. *Annu. Rev. Microbiol.* **61**, 191–214 (2007).
- 671 19. Sperandio, P., Martorana, A. M. & Polissi, A. The Lpt ABC transporter for lipopolysaccharide  
672 export to the cell surface. *Res. Microbiol.* (2019) doi:10.1016/j.resmic.2019.07.005.
- 673 20. Heinz, E., Selkrig, J., Belousoff, M. J. & Lithgow, T. Evolution of the translocation and  
674 assembly module (TAM). *Genome Biol. Evol.* **7**, 1628–1643 (2015).
- 675 21. Noinaj, N., Guillier, M., Barnard, T. J. & Buchanan, S. K. TonB-Dependent Transporters:  
676 Regulation, Structure, and Function. *Annu. Rev. Microbiol.* **64**, 43–60 (2010).
- 677 22. Hughes, G. W. *et al.* Evidence for phospholipid export from the bacterial inner membrane by  
678 the Mla ABC transport system. *Nat. Microbiol.* (2019) doi:10.1038/s41564-019-0481-y.
- 679 23. Mukherjee, S. & Kearns, D. B. The Structure and Regulation of Flagella in *Bacillus subtilis* .  
680 *Annu. Rev. Genet.* **48**, 319–340 (2014).
- 681 24. Jacquier, N., Yadav, A. K., Pilonel, T., Viollier, P. H. & Greub, G. A SpoIID Homolog  
682 Cleaves Glycan Strands at the Chlamydial Division Septum. *Mol. Biol. Physiol.* **10**, 1–16 (2019).
- 683 25. Delsuc, F., Brinkmann, H. & Philippe, H. Phylogenomics and the reconstruction of the tree of  
684 life. *Nat. Rev. Genet.* **6**, 361–375 (2005).
- 685 26. Cavalier-Smith, T. Rooting the tree of life by transition analyses. *Biol. Direct* **1**, 1–83 (2006).
- 686 27. Battistuzzi, F. U. & Hedges, S. B. A major clade of prokaryotes with ancient adaptations to life  
687 on land. *Mol. Biol. Evol.* **26**, 335–343 (2009).
- 688 28. Lake, J. A. Evidence for an early prokaryotic endosymbiosis. *Nature* **460**, 967–971 (2009).



- 689 29. Vollmer, W. Bacterial outer membrane evolution via sporulation? *Nat. Chem. Biol.* **8**, 14–18  
690 (2012).
- 691 30. Tocheva, E. I. *et al.* Peptidoglycan remodeling and conversion of an inner membrane into an  
692 outer membrane during sporulation. *Cell* **146**, 799–812 (2011).
- 693 31. Sutcliffe, I. C. A phylum level perspective on bacterial cell envelope architecture. *Trends*  
694 *Microbiol.* **18**, 464–470 (2010).
- 695 32. Raymann, K., Brochier-Armanet, C. & Gribaldo, S. The two-domain tree of life is linked to a  
696 new root for the Archaea. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6670–6675 (2015).
- 697 33. Cavalier-Smith, T. & Chao, E. E. Y. Multidomain ribosomal protein trees and the  
698 planctobacterial origin of neomura (eukaryotes, archaebacteria). *Protoplasma* 621–753 (2020)  
699 doi:10.1007/s00709-019-01442-7.
- 700 34. Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1**, 1–6 (2016).
- 701 35. Vincent, A. T. *et al.* The mycobacterial cell envelope: A relict from the past or the result of  
702 recent evolution? *Front. Microbiol.* **9**, 1–9 (2018).
- 703 36. Gaisin, V. A., Kooger, R., Grouzdev, D. S., Gorlenko, V. M. & Pilhofer, M. Cryo-Electron  
704 Tomography Reveals the Complex Ultrastructural Organization of Multicellular Filamentous  
705 Chloroflexota (Chloroflexi) Bacteria. *Front. Microbiol.* **11**, 1–15 (2020).
- 706 37. Sutcliffe, I. C. Cell envelope architecture in the Chloroflexi: A shifting frontline in a  
707 phylogenetic turf war. *Environ. Microbiol.* **13**, 279–282 (2011).
- 708 38. Blobel, G. Intracellular protein topogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **77**, 1496–500 (1980).
- 709 39. Hyatt, D. *et al.* Prodigal: Prokaryotic gene recognition and translation initiation site  
710 identification. *BMC Bioinformatics* **11**, (2010).
- 711 40. Finn, R. D. *et al.* The Pfam protein families database: Towards a more sustainable future.  
712 *Nucleic Acids Res.* **44**, D279–D285 (2016).
- 713 41. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative  
714 HMM search procedure. *BMC Bioinformatics* **11**, (2010).
- 715 42. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: A new generation of protein database  
716 search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
- 717 43. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments  
718 using Clustal Omega. *Mol. Syst. Biol.* **7**, (2011).
- 719 44. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new  
720 software for selection of phylogenetic informative regions from multiple sequence alignments.  
721 *BMC Evol. Biol.* **10**, 210 (2010).

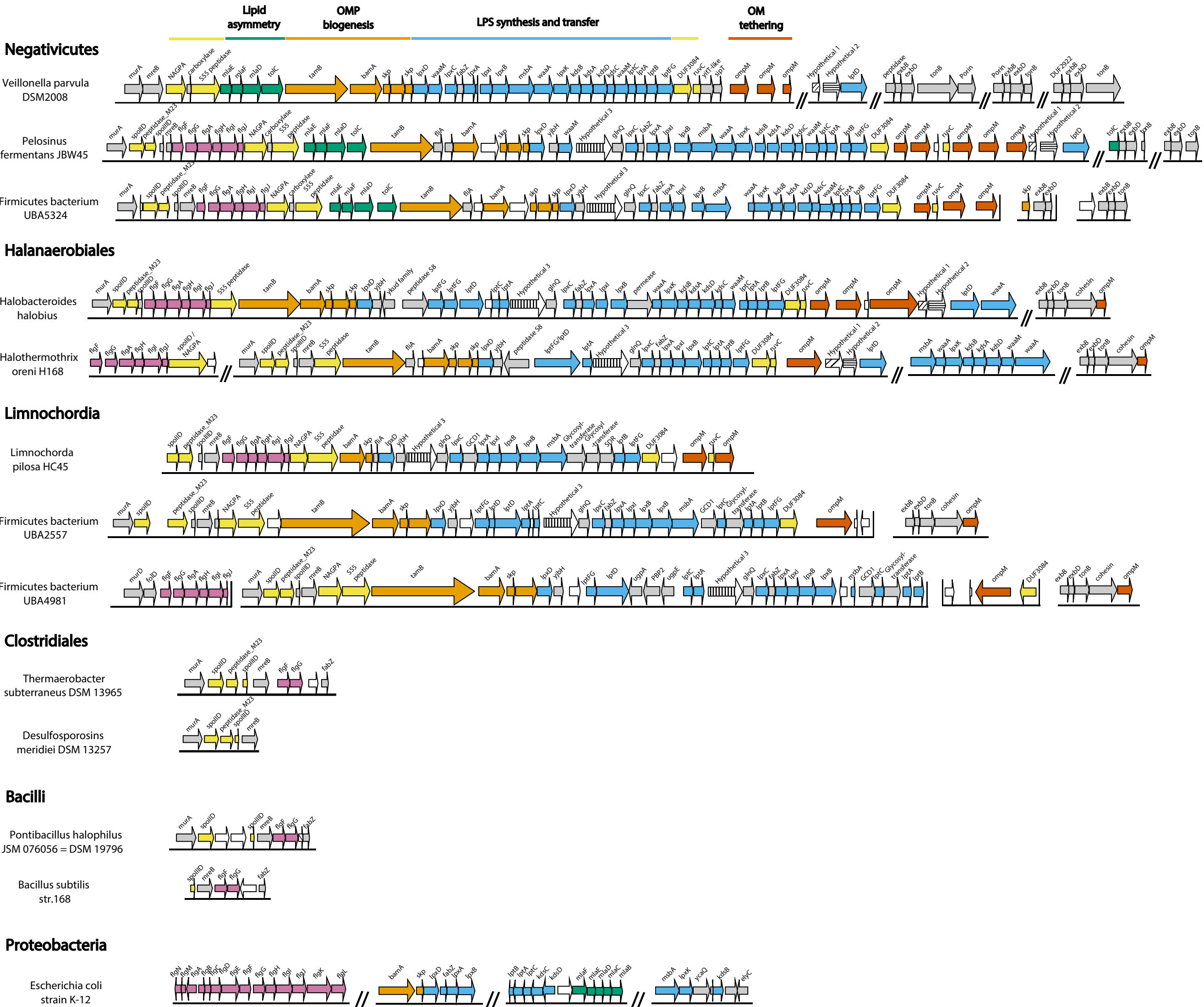
- 722 45. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and  
723 effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*  
724 **32**, 268–274 (2015).
- 725 46. Minh, B. Q., Nguyen, M. A. T. & Von Haeseler, A. Ultrafast approximation for phylogenetic  
726 bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
- 727 47. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the  
728 amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
- 729 48. Letunic, I. & Bork, P. Interactive Tree of Life v2: Online annotation and display of  
730 phylogenetic trees made easy. *Nucleic Acids Res.* **39**, 1–4 (2011).
- 731 49. Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment  
732 by eggNOG-mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
- 733 50. Huerta-Cepas, J. *et al.* EGGNOG 4.5: A hierarchical orthology framework with improved  
734 functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**,  
735 D286–D293 (2016).
- 736 51. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with  
737 SiLiX. *BMC Bioinformatics* **12**, (2011).
- 738 52. Miele, V. *et al.* High-quality sequence clustering guided by network topology and multiple  
739 alignment likelihood. *Bioinformatics* **28**, 1078–1085 (2012).
- 740 53. Garcia, P. S., Jauffrit, F., Grangeasse, C. & Brochier-Armanet, C. GeneSpy, a user-friendly  
741 and flexible genomic context visualizer. *Bioinformatics* **35**, 329–331 (2019).
- 742 54. Abby, S. S., Néron, B., Ménager, H., Touchon, M. & Rocha, E. P. C. MacSyFinder: A  
743 program to mine genomes for molecular systems with an application to CRISPR-Cas systems.  
744 *PLoS One* **9**, (2014).
- 745 55. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:  
746 Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
- 747 56. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new  
748 developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
- 749





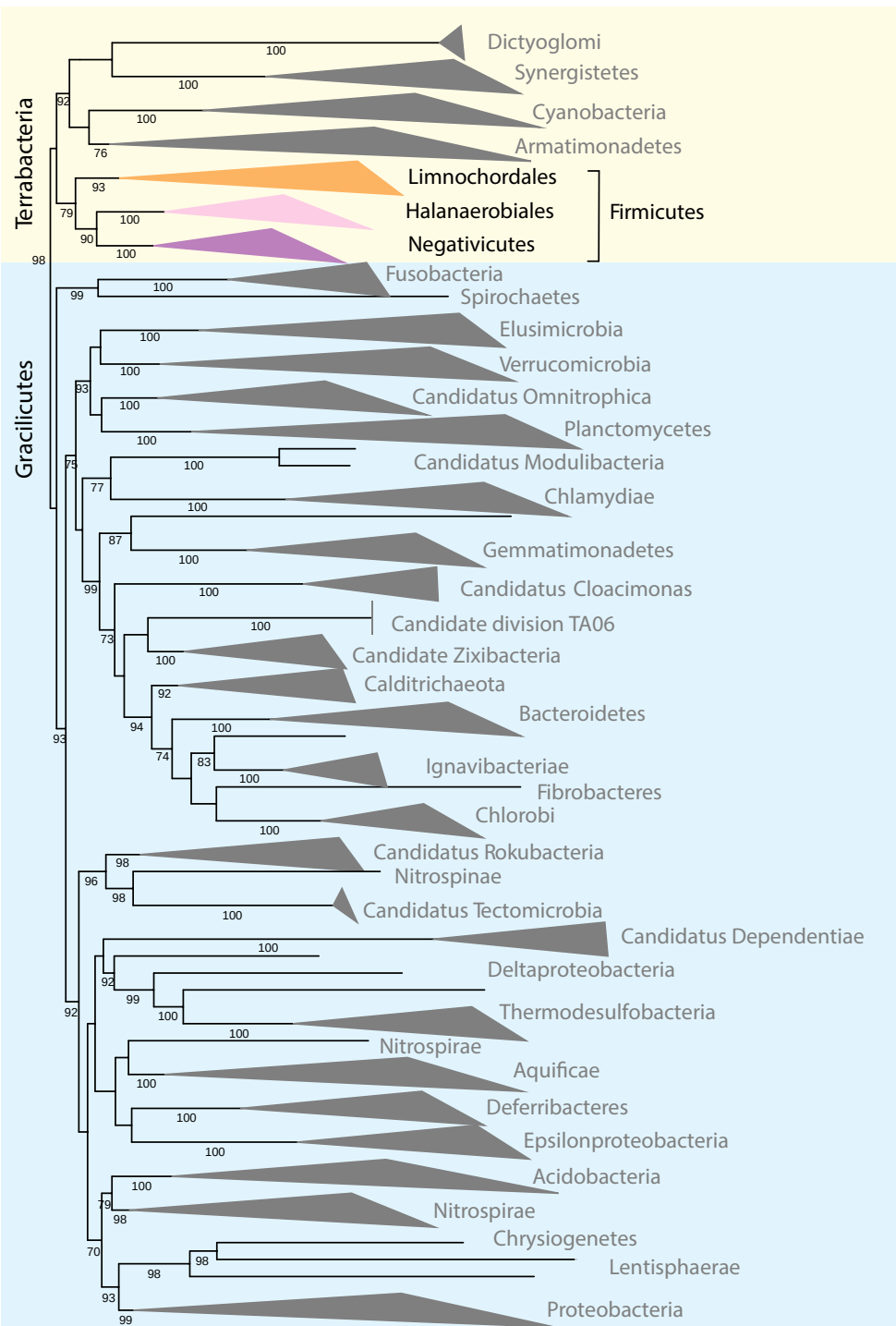
Protein Family/ Pfam Domain	Correlation coefficient	Diderm (165)	Monoderm (151)	Pfam Annotation	Putative Function	OM cluster
PF06835	0.97	159	0	LptC	LptC (LPS transport)	YES
PF01103	0.96	161	0	Bac_surface_Ag	BamA (OMP transport and insertion)	YES
PF11283	0.96	160	1	DUF3084	uncharacterized	YES
PF03968	0.92	155	2	OstA	LptA, LptD (LPS transport)	YES
FAM35001205_1	0.92	153	0	RuvX	RuvC (endonuclease)	YES
PF07660,PF03958,PF00263	0.92	157	3	STN,Secretin_N,Secretin	PilQ (Type II secretion)	
PF02684	0.91	152	2	LpxB	LpxB (LPS synthesis)	YES
FAM35003477_0	0.91	151	0	YjgP_YjgQ	LptFG (LPS transport)	YES
PF04357	0.91	151	0	TamB	TamB <sup>1</sup> (autotransporter assembly complex)	YES
FAM35003763_0	0.9	150	0	MotA_ExbB	ExbB (energy transducing Ton complex)	
PF03938	0.89	163	22	OmpH	Skp (OMP transport and insertion)	YES
PF02472	0.87	152	6	ExbD	ExbD (energy transducing Ton complex)	
FAM35001011_0	0.85	141	0	Peptidase_S55	Peptidase	YES
PF13502	0.85	141	2	AsmA_2	TamB <sup>1</sup> (autotransporter assembly complex)	YES
PF03544	0.84	140	1	TonB_C	TonB (energy transducing Ton complex)	
FAM35006918_0	0.83	137	0	Hexapep,Acetyltransf_11	LpxA (LPS synthesis)	YES
FAM35003933_0	0.83	138	0	Hexapep,LpxD	LpxD (LPS synthesis)	YES
FAM35004424_0	0.82	135	0	DUF1009	LpxI (LPS synthesis)	YES
PF03331	0.82	140	3	LpxC	LpxC (LPS synthesis)	YES
PF13505	0.79	131	1	OMP_b-brl	OM beta barrel	
FAM35002111_0	0.77	122	0	Lip_A_acyltrans	WaaM (LPS synthesis)	YES
FAM35000110_0	0.76	123	0	LpxK	LpxK (LPS synthesis)	YES
FAM35001087_0	0.76	123	0	TrbI	hypothetical 2 (Bacterial conjugation TrbI-like protein)	YES
FAM35002134_0	0.75	122	0	Hydrolase_3	KdsC (LPS synthesis)	YES
FAM35001758_0	0.75	124	1	CTP_transf_3	KdsB (LPS synthesis)	YES
PF04413	0.74	125	3	Glycos_transf_N	waaA (LPS synthesis)	YES
FAM35007211_0	0.74	121	1	DAH_p_synth_1	KdsA (LPS synthesis)	YES
FAM35001331_0	0.73	119	1	Glyco_transf_9	glycosyl transferase	
FAM35000503_0	0.73	121	2	SIS,CBS	KdsD (LPS synthesis)	YES
PF05258	0.71	114	0	TonB_dep_Rec	TonB complex, OM receptor	
PF00593	0.71	126	8	DUF721	uncharacterized	
FAM35000108_2	0.66	138	25	Peptidase_M23	peptidase	YES
PF11741	0.65	145	34	AMIN	AmC N-terminal domain	
PF11209	0.62	104	8	DUF2993	uncharacterized	
FAM35007523_0	0.61	123	21	RecX	modulates RecA activity	
FAM35012460	0.59	90	0	FlgH	FlgH (flagellum L-ring)	YES
FAM35002503	0.59	90	0	FlgI	FlgI (flagellum P-ring)	YES
PF00395	0.59	165	74	SLH	OmpM <sup>2</sup> (OM tethering)	YES
FAM35001613	0.58	88	0	Sigma70_r2,Sigma70_r4,Sigma70_r1_2	Transcription factor	
FAM35001477_0	0.56	152	58	Prbonyltran	Phosphonobosyl transferase domain	
FAM35002373_0	0.56	140	44	PS_pyruv_trans	Polysaccharide pyruvyl transferase	
PF02563,PF10531	0.55	78	0	Poly_export,SLBB	Polysaccharide biosynthesis/export protein, ligand binding	
FAM35000742_0	0.55	121	28	Germane	GerM (spore germination/cell division)	
PF04402	0.55	136	41	SIMP1	uncharacterized	
FAM35000495_0	0.54	139	46	DwIC	Septum formation	
PF13609	0.53	74	0	Porin_4	OmpM <sup>2</sup> (OM tethering)	YES
PF03797	0.52	73	1	Autotransporter	Type V(a) secretion	
FAM35006277	0.51	70	0	SurA_N_3	Outer membrane transporter	
PF03865	0.51	75	1	ShIB	Type V(b) secretion	
FAM35005930_0	0.51	155	72	B12-binding,Radical_SAM	unclear	
FAM35008032_0	0.51	115	28	Om_Arg_deC_N	hypothetical 1 (Om/Lys/Arg decarboxylase class-II family)	YES
FAM35004850_0	0.5	104	20	LysM	PG binding	

Protein Family/ Pfam Domain	Correlation coefficient	Diderm (165)	Monoderm (151)	Pfam Annotation	Putative Function	OM cluster
FAM35000114_0	0.94	0	142	S4 (RNA-binding, translation regulation)		
PF04203.12	0.72	1	104	Sortase (attaches surface proteins to the cell wall)		
FAM35004194_0	0.7	21	126	THUMP,UPF0020 (putative RNA methylase)		
FAM35002871_0	0.67	44	140	CDP-OH_P_transf (CDP-alcohol phosphatidyltransferase)		
FAM35000917_0	0.67	18	117	PLDc_N,PLDc_2 (phospholipase)		
PF01170.17	0.66	36	133	UPF0020 (Putative RNA methylase)		
PF08353.9	0.64	0	87	DUF1727 (C-terminus of bacterial proteins which include UDP-N-acetylmuramyl tripeptide synthase and the related Mur ligase)		
PF07261.10	0.64	54	144	DnaB_2 (replication initiation and membrane attachment)		
PF05389.11	0.64	0	86	MecA (negative regulator of competence)		
PF03951.18	0.63	32	126	Gln_synth_N (glutamine synthetase)		
FAM35005306_0	0.6	19	107	Hpr_kinase_N,Hpr_kinase_C (serine kinase)		
PF00746.20	0.59	20	105	Gram_pos_anchor (LPXTG cell wall anchor motif)		
FAM35000755	0.59	0	77	GntR,4HBT,CBS,DRTGG (unclear)		
PF08245.11,PF08353.9	0.58	0	76	Mur_ligase_M,DUF1727 (Mur ligase middle domain)		
PF01521.19	0.58	4	83	Fe-S_biosyn (iron-sulfur cluster biosynthesis)		
PF02325.16	0.58	37	123	YGGT (repeat found in conserved hypothetical integral membrane proteins)		
FAM35002460	0.57	0	74	GATase_3 (CobB/CobQ-like glutamine amidotransferase domain)		
FAM35005063	0.57	4	80	Lactamase_B_2 (Beta-lactamase superfamily domain)		
PF12679.6	0.57	45	126	ABC2_membrane_2 (ABC-2 family transporter protein)		
FAM35001568	0.56	0	72	NAD_kinase (ATP-NAD kinase)		
FAM35001032_0	0.56	21	101	Toprim,Topoisom_bac,Toprim_Crp1 (topoisomerase)		
PF01694.21	0.56	46	126	Rhomboid (intramembrane protease)		
PF02517.15	0.56	76	148	Abi (abortive infection phage resistance)		
PF06103.10	0.56	19	100	DUF948 (unknown function)		
FAM35002806_1	0.56	31	113	PGI (phosphoglucose isomerase)		
PF09648.9	0.55	3	75	YycI (regulate the essential YycFG two-component system in Bacillus subtilis)		
PF04167.12	0.54	0	68	DUF402 (unknown function)		
FAM35008547	0.54	0	68	Prbonyltran (Phosphonobosyl transferase domain)		
FAM35004009_0	0.54	4	75	Radical_SAM,Radical_SAM_C (Radical_SAM)		
FAM35003839	0.54	2	71	MazG (nucleotide pyrophosphohydrolase)		
FAM35002605	0.54	0	67	Wzz_GNVR (O-antigen chain length, G-rich domain on putative tyrosine kinase)		
PF07435.10	0.54	0	67	YycH (regulates the essential YycFG two-component system in Bacillus subtilis)		
FAM35001229_1	0.53	65	138	PGM_PMM_I,PGM_PMM_II,PGM_PMM_III,PGM_PMM_IV		
PF06207.10	0.53	4	73	DUF1002		
PF07319.10	0.53	1	67	DnaI_N (Primosomal protein DnaI N-terminus)		
PF01883.18	0.52	3	70	FeS_assembly_P (Iron-sulfur cluster assembly protein)		
FAM35003271_0	0.52	47	123	TPK_catalytic,TPK_B1_binding Thiamin pyrophosphokinase)		
FAM35005240_0	0.52	17	91	PNP_UDP_1 (Phosphorylase superfamily)		
FAM35002295_0	0.52	4	71	Penicillinase_R (Penicillinase repressor)		
PF01145.24	0.52	40	116	Band_7 (slipin or Stomatol-like integral membrane domain)		
PF01643.16	0.52	2	67	Acyl-ACP_TE (Acyl-ACP thioesterase)		
PF00355.25	0.52	36	74	Rieske (Rieske [2Fe-2S] domain)		
PF03631.14	0.52	26	101	Virul_fac_BrKb (Virulence factor BrkB)		
PF05975.11	0.51	7	75	EcsB (ABC transporter)		
FAM35000288_1	0.51	1	64	Pyr_redox_2 (Pyridine nucleotide-disulphide oxidoreductase)		
FAM35002436_1	0.51	9	78	AMP-binding,AMP-binding_C (AMP-binding enzyme)		
PF06081.10	0.51	24	96	ArAE_1 (Aromatic acid exporter family member 1)		
PF02687.20,PF02687.20	0.5	18	87	FtsX,FtsX (FtsX-like peptidase family)		
FAM35006543	0.5	0	60	MMR_HSR1 (50S ribosome-binding GTPase)		
PF04026.11	0.5	46	119	SpoVG (Stage V sporulation protein G)		
PF15980.4	0.5	1	62	ComGF (Putative Competence protein ComGF)		

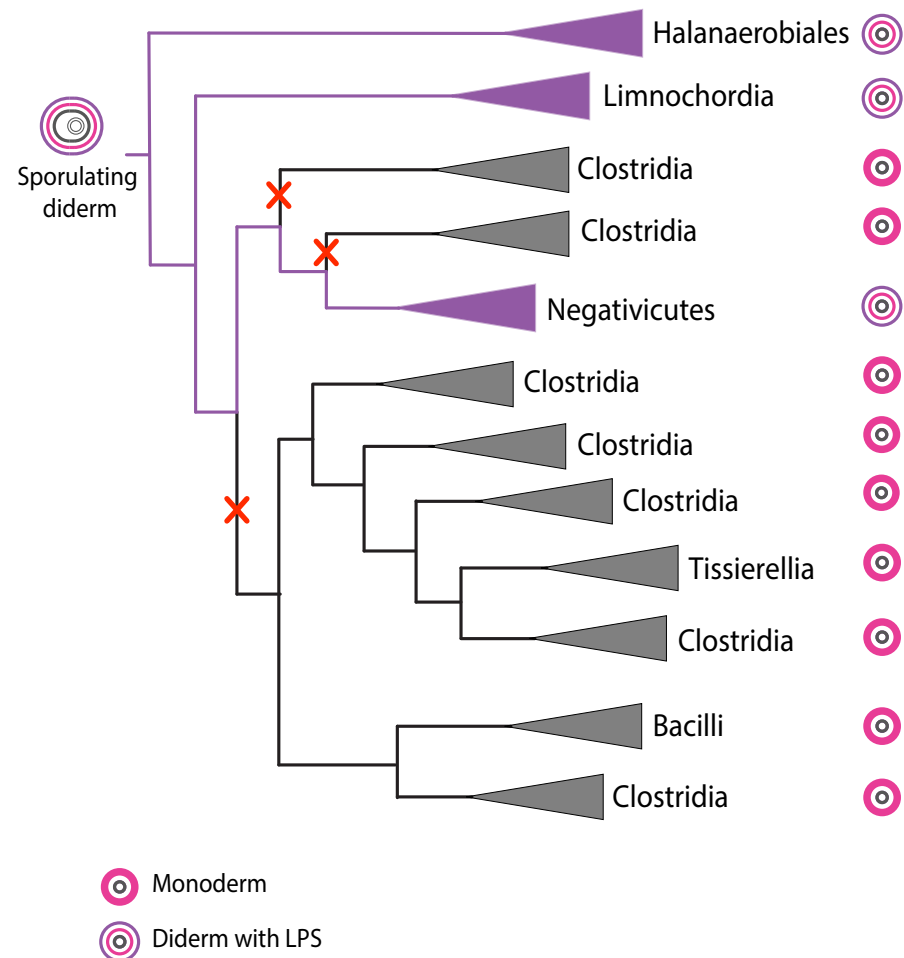


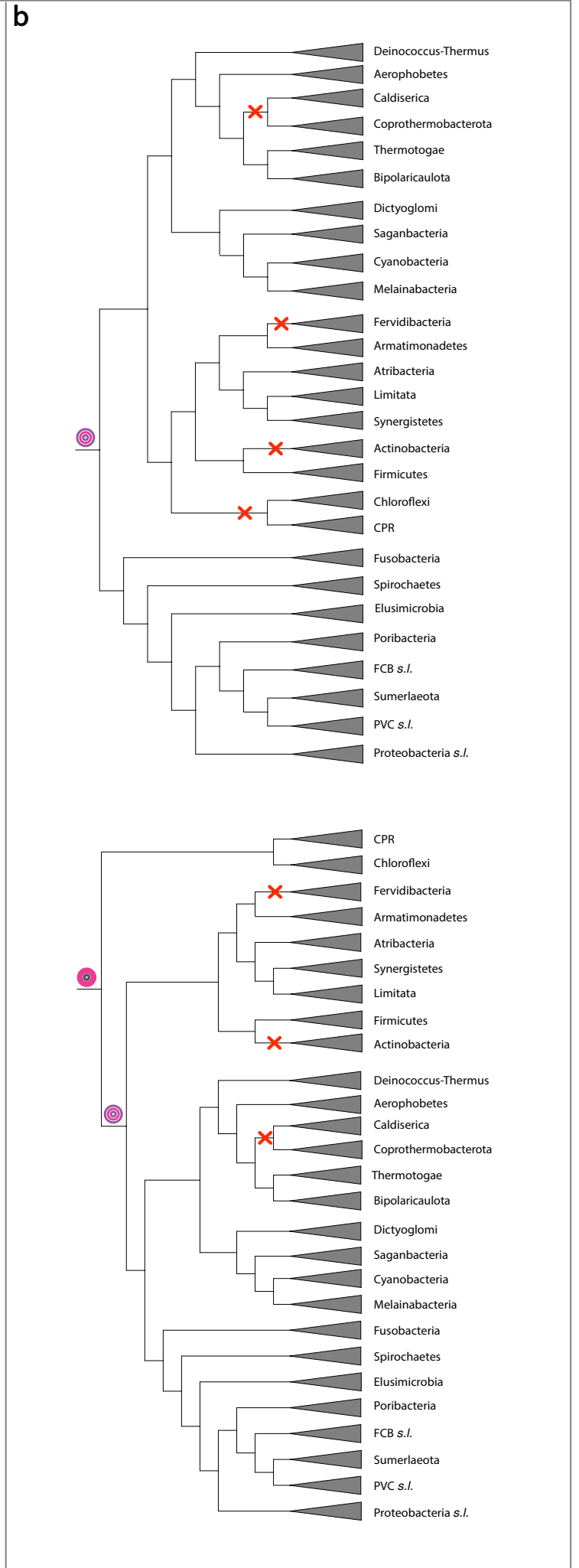
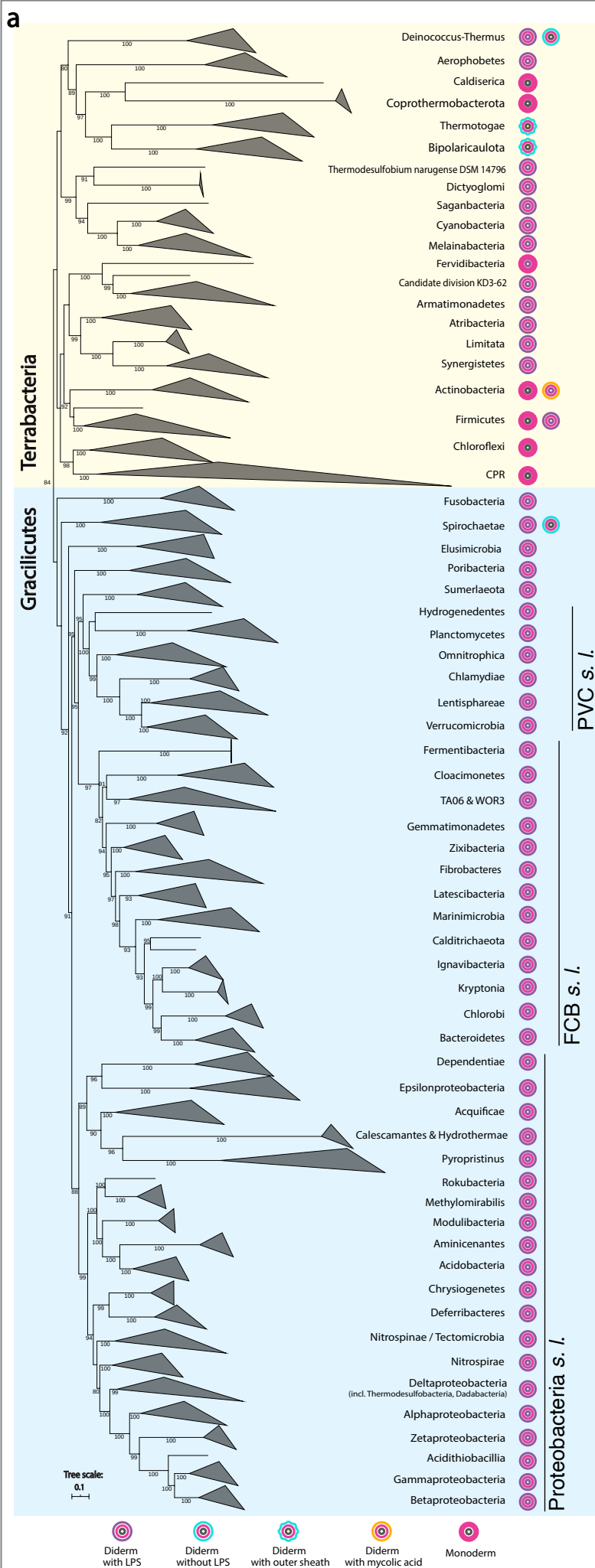


a



b





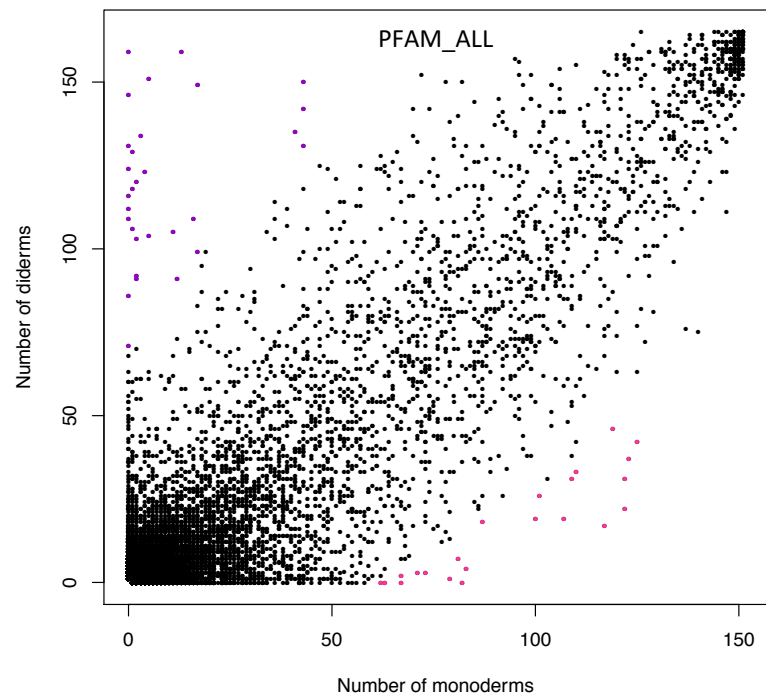
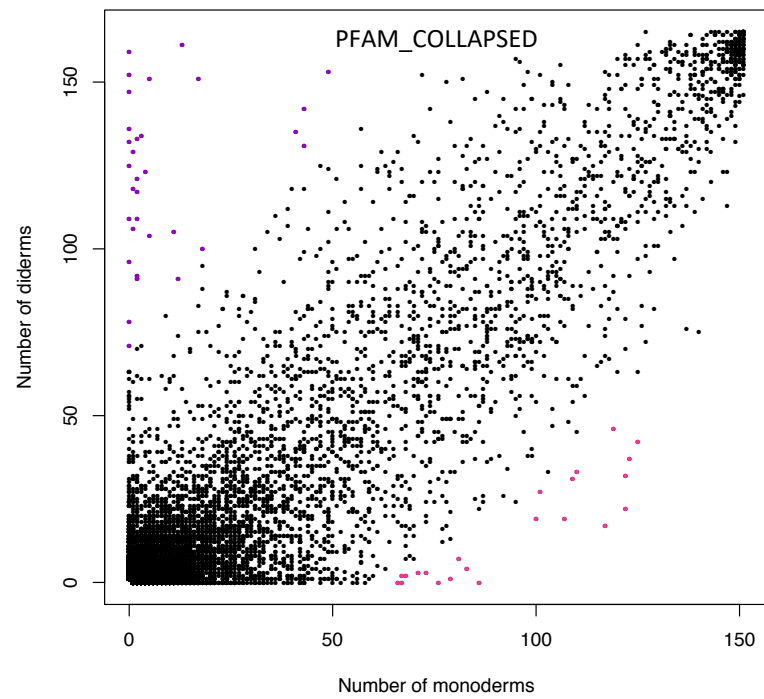
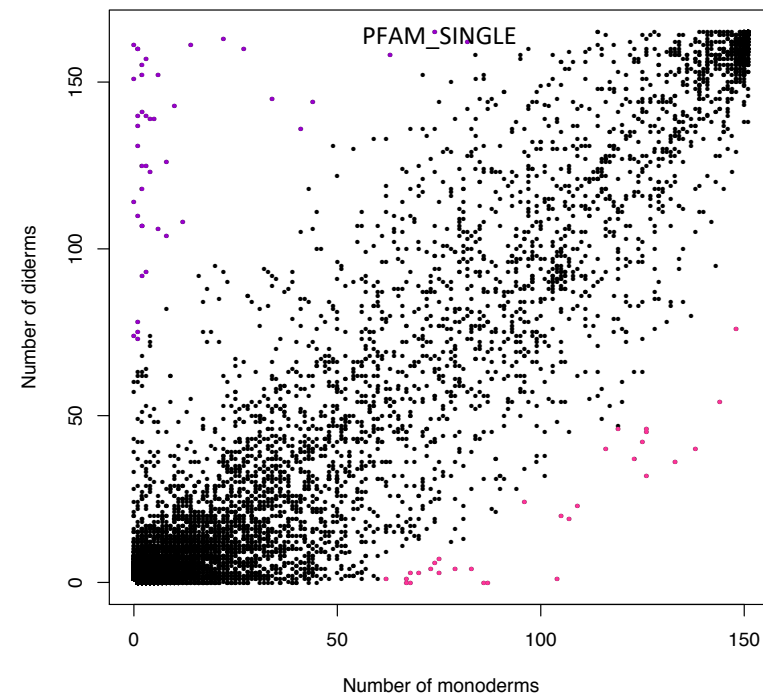


Tree scale: 0.1



**Supplementary figure 1: Phylogeny of the Firmicutes based on a reduced taxon sampling corresponding to the genomes used for the construction of the protein families. Related to Figure 1.**

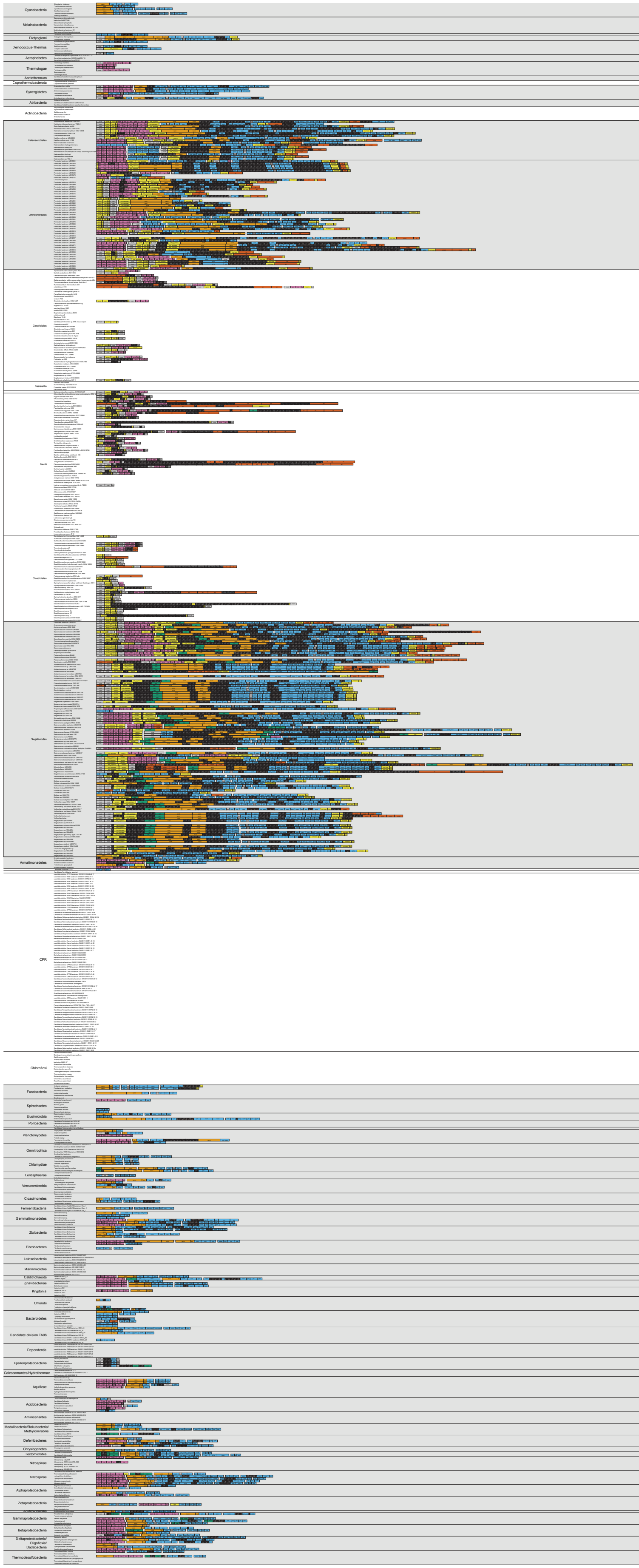
Maximum likelihood tree based on concatenation of 45 ribosomal proteins (329 taxa, 5,382 amino acid characters), inferred with IQ-TREE 1.4.4 using the LG+C60+F+G model. Dots at nodes represent bootstrap values (BV) higher than 80% calculated on 100 replicates of the original dataset. The scale bar corresponds to the average number of substitutions per site. Spo0A was mapped on this tree by scanning the taxa for the presence of Spo0A\_C (PF08769.11) using HMMSEARCH. Red bars indicate the presence of spo0A.

**A.****B.****C.**

**Supplementary figure 2: Distribution of PFAM domains according to their presence in 165 diderm and 151 monoderm Firmicutes. Related to Figure 3.**

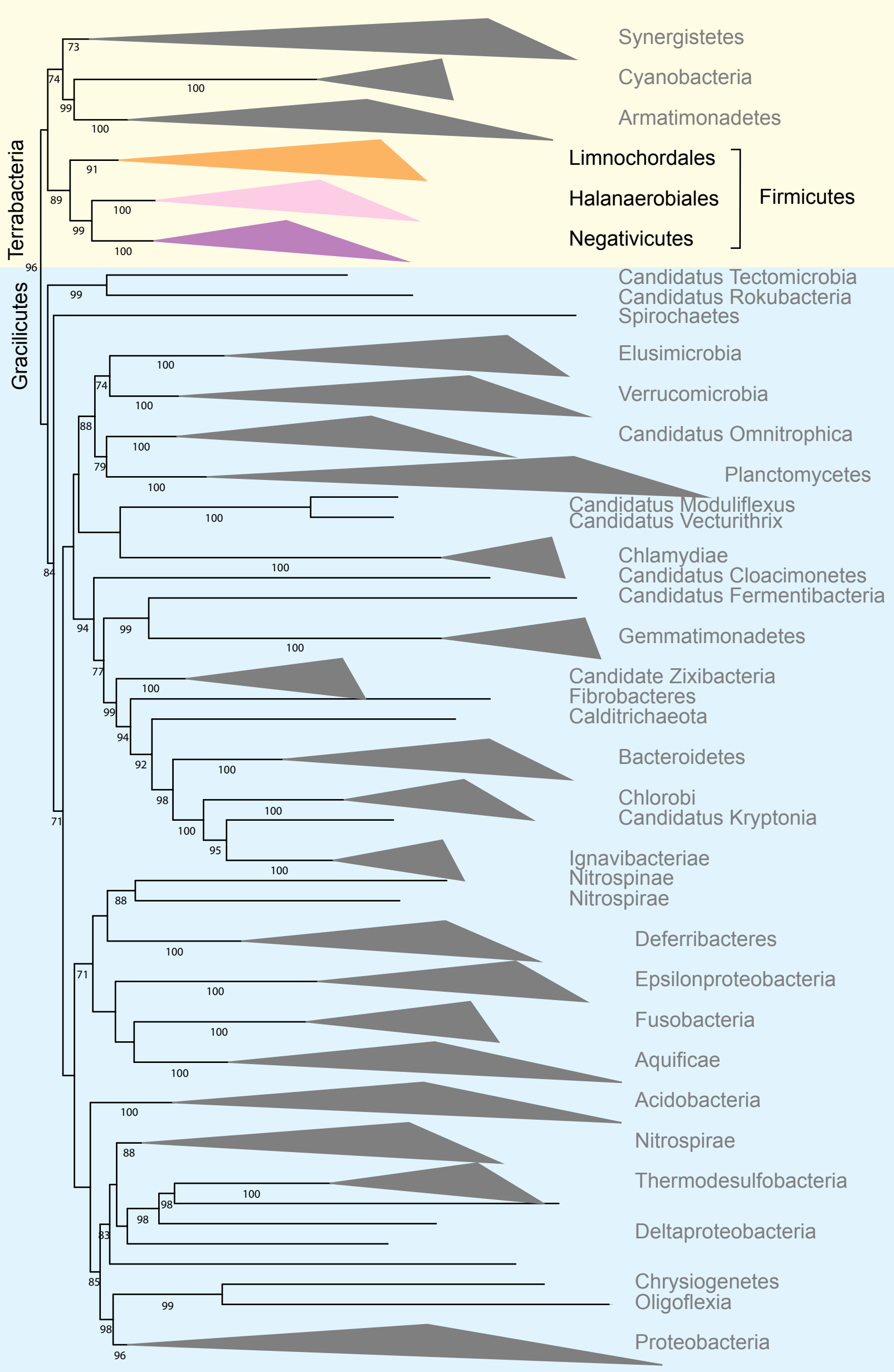
Dots in purple and pink correspond to domains with a correlation higher than 0.5 with the presence or absence of an OM respectively. The three approaches are represented: PFAM\_ALL, PFAM\_COLLAPSED and PFAM\_SINGLE.

- (a) Distribution of 56,513 PFAM\_ALL domains.
- (b) Distribution of 51,413 PFAM\_COLLAPSED domains.
- (c) Distribution of 14,808 PFAM\_SINGLE domains.



**Supplementary figure 3: Organization of OM clusters in Bacteria. Related to Figure 4.**

OM clusters are listed for all taxa in DB BACTERIA and DB SMALL Firmicutes. Phyla in gray are diderm, those in white are monoderm. Color codes are as in Figure 5.



Tree scale: 0.1

**Supplementary figure 4: Phylogeny of OM cluster components including flagellar proteins. Related to Figure 5.**

Maximum likelihood tree based on concatenation of 17 proteins (FabZ, KdsA, KdsB, KdsC, KdsD, LptB, LpxA, LpxB, LpxC, LpxD, LpxK, FlgF, FlgG, FlgA, FlgH, FlgI and FlgJ) comprising 122 taxa and 3,705 amino acid positions, inferred with IQ-TREE 1.4.4 using the model LG+C60+F+G and ultrafast bootstrap (UFB) supports calculated on 1000 replicates of the original dataset. UFB higher than 70% are displayed. The scale bar corresponds to the average number of substitutions per site.